DOCUMENT RESUME

ED 193 263 TM 800 577

AUTHOR	Wood, Peter H.; And Others
TITLE	Student Ratings of College Courses and College Teachers: Reliability, Generalizability, Validity and Utility.
INSTITUTION	Bowling Green State Univ., Ohio. Coll. of Education.
PUB DATE	80
NOTE	172p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Behavior Rating Scales; Faculty Development; Higher Education; *Reliability; *Student Evaluation of Teacher Performance; *Testing Problems; *Test Validity; *Use Studies

ABSTRACT

A review of the research literature considers the following questions: (1) what is the general quality of the information resulting from systematic student evaluation of faculty effectiveness?; and (2) what are the consequences of soliciting this information and making it available to those people who may utilize it in their decision making procedures? The internal consistency of most ratings scales, and many subscales, often exceed .9, and, when lower, often can be increased by increasing the number of ratings items in a particular scale. However, this indicator of reliability produces a considerable overestimation of ratings stability when independent measures of teachers and/or courses are desired. Generalizability theory has begun to be employed to supplement and to extend the more traditional approaches to reliability. Factors such as expected grades, class-size, and the students' pre-enrollment attitudes are likely to influence student ratings. Much of the instability and potential invalidity may be reduced through the standardization of administrative procedures and the use of global rating items for normative or personnel purposes. There appears, however, to be little evidence that student ratings are weighted very heavily in the decision-making processes of personnel committees at the college level. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193263

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Student Ratings of College Courses and College Teachers: Reliability, Generalizability, Validity and Utility

Peter H. Wood
Genevieve Stang
Jane H. Wood

College of Education
Bowling Green State University
Bowling Green, Ohio 43403

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Wood, Stang,
Wood

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TM 800577

ABSTRACT

This study represents a review of the research literature describing student ratings of college courses and the teaching occurring in these courses. Students are major beneficiaries of effective instruction, bear many of the expenses associated with ineffective instruction, and are in a unique position to evaluate or comment upon many aspects of classroom instruction. Although the ratings of classes of twenty or more students are at least moderately reliable, and often as reliable as scores produced by professionally developed achievement tests, too many researchers have failed to report the types of reliability or generalizability characteristics that are most appropriate to the different functions for which ratings data are employed. The more recent employment of generalizability theory and its associated procedures have tended to correct the frequent misuse or misinterpretation of the reliability data descriptive of student ratings.

A similar failure to examine ratings results in terms of the different uses of the ratings data, or the actual questions asked of students, has tended to cloud many of the attempts to examine the validity of this type of student self-report. For example, student reports of their current grade point average, of their expected grade in a course, or of their estimates of the academic ability of their fellow students all appear to be as valid as the students' degree of current knowledge permits them to be. Student reports of their degree of learning in a course also tend to correlate as well with other indicators

of student learning as these other indicators correlate with one another. Student reports of their satisfaction with a course or with an instructor tend to be less correlated with external indicators of student learning while often correlated with expected course grade or with other likely predictors of satisfaction such as class size, degree of positive interaction, and previous interest in the subject. Altogether too often, evidences of validity have been confused with evidences of invalidity due to a lack of clarity concerning the real meaning of the different questions asked of students by designers of rating forms.

The most crucial questions, those concerning the actual effects that ratings systems have upon users, rarely have been studied. There is limited evidence that ratings of college courses and instruction actually lead to much subsequent improvement unless additional supportive or developmental systems are employed in addition to ratings. There is little or no evidence as to the effects of ratings use upon personnel practices, faculty or student satisfaction, collegiality, instructional efficiency, student or faculty self-esteem, or student appreciation or understanding of the instructional process.

Student Ratings of College Courses and College Teachers: Reliability, Generalizability, Validity and Utility

Peter H. Wood
Genevieve Stang
Jane H. Wood

College of Education
Bowling Green State University
Bowling Green, Ohio 43403

THE INTRODUCTION

Students are major beneficiaries of effective teaching, and students bear many of the expenses associated with ineffective instruction. As the primary "consumer" and the most frequent observer of classroom instruction, the college student is in a unique position to comment upon faculty performance and instructional effectiveness. However, the student, although the only person who is exposed to the full range of a teacher's classroom performance, also tends to be one of the youngest, least mature, least experienced, least powerful, and least statusful member of the college community. The major questions relating to student evaluation of faculty effectiveness do not ask whether or not a student should or should not evaluate or judge teaching effectiveness, since it would be virtually impossible to prevent either student judgment or student communication of their judgments to other members of the college community. The major questions are:

1. What is the general quality of the information resulting from systematic student evaluation of faculty effectiveness (reliability, generalizability, validity); and

2. What are the consequences of soliciting this information and making it available to those people who may utilize it in their decision making procedures (utility)?

Recent History

Less than ten years ago, Costin, Greenough, and Menges (1971) published a comprehensive review of student ratings of college teaching and concluded that "a review of empirical studies indicates that students' ratings can provide reliable and valid information on the quality of courses and instruction.... (and) such information can be of use to academic departments in constructing normative data for the evaluation of teaching and may aid the individual instructor in improving his teaching effectiveness" (p. 530). These authors also suggested that perceived threat to faculty may have inhibited the more extensive employment of student ratings, "since in recent years a strong impetus to use students ratings has come from the students themselves" (p. 511). Since 1971, there has been a considerable growth in the body of research which describes student ratings and faculty evaluation. Some of the earlier conclusions have gained additional support, other conclusions have been compromised by more recent research, and new avenues for study have been opened to exploration. Although much of the research concerning student ratings and faculty evaluation has proceeded in predictable directions, the academic environment within which this evaluation and its research has operated also has changed in many ways, several of which were only predicted by the more foresighted.

1. Faculty mobility has sharply decreased since 1971 (Cartter, 1976, p. 161).
2. Active enforcement of federal and state anti-bias regulations (Lerner, 1978a, 1978b; Lester, 1974; Novick, 1978) has been a noticeable modifier of academic, personnel procedures.

3. The increasing rigor of promotion and tenure procedures has been documented by Shull (1976).
4. Staff reductions, especially noticeable in Colleges of Education, have created an entirely new area of expertise and a concomitant opportunity for professional writing which is evident in these recent article titles: "Legal Issues in Faculty Termination: An Analysis Based on Recent Court Cases" (Rood, 1977); "Managing Faculty Reductions" (Alm, Ehrle, & Webster, 1977); and "Tenure: A Sacred Cow?" (Mayr, 1978).

Each of these changes has created an increased need to clarify and validate those procedures used to evaluate faculty; each change has increased institutional needs to provide development and support services which usually include faculty evaluation; and each change has provided many faculty with additional threats to security which are likely to exceed those threats produced by dissident students of the "turbulent 60's". The national movements toward accountability, economic factors which diminish faculty buying power, tax revolts which threaten the financial supports of the educational system, and the increasingly union-like activity by college faculties—all also represent trends which can lead toward greater efforts to evaluate faculty while also increasing the level of threat to faculty security. Sedin and Wakin's (1974) book provided evidence that deans were critical of their current evaluation procedures and were worried that these procedures were not based upon "rational, impersonal, and unprejudiced information" (p. 49). Comparison of Sedin and Wakin's data with data produced by Astin and Lee in 1966 indicates the trend toward the increased employment of more structured and more systematic faculty evaluation procedures: the increased use of formal student ratings has surpassed the less formal gathering of student opinions; and indicators of research activities rather than indicators of the teaching function have fallen into disuse as indirect measures of teaching effectiveness. Centra's survey of heads of university and college departments "also suggest an increase in the use of some of the more systematic and tangible kinds of evidence for evaluating teaching performance (e.g., student ratings), though the use of such evidence falls far short of what department heads think is needed" (Centra, 1977c, p. 19). Similar movement toward more structured and rational evaluation of faculty is also evident in Sedin and Wakin's findings that formal committee evaluation of faculty performance has increased at the expense of informally obtained, collegial opinions. Shipka (1978) has warned that peer review systems will face a series of court tests unless peer committees systematically and regularly acquire reliable information about the quality of faculty performance in teaching, scholarship, and college service, and such threats may speed the movement toward more formal evaluation of faculty performance.

Student Ratings: Current States

Centra's (1976) survey of American degree-granting schools led him to state: "half, or slightly over half, of the post-secondary institutions in the United States now provide some sort of program or set of staff development activities" (Centra, 1977a, p. 50-51); that "systematic ratings by students to help faculty to improve instruction are used at more than 80% of the institutions" (1977b, p. 104); that "the recent emphasis on faculty development and instructional improvement programs is in sharp contrast to what the majority of colleges were doing just a few years ago" (p. 49)--the median number of years of program existence was only 2.3 years (Table 10, p. 34); and that "the increased use of student ratings over the past decade seems rivaled only by the growing numbers of articles that condemn or support them" (p. 95).

Carefully constructed student ratings forms have long been available from those colleges that pioneered in their use. The various University of Washington forms and data describing them have been available since 1924 (Guthrie, 1949, 1954). Remmers and his associates at Purdue (Remmers and Brandenburg, 1927) published descriptions of their form a few years later. McKeachie's Teaching Tips: A Guidebook for the Beginning College Teacher (1965) includes his excellent ratings form as well as suggestions for the use of student ratings by college teachers, and the first of many issues of his guide appeared in 1951. The Course Evaluation Questionnaire (CEQ) was developed at the University of Illinois and was reported in 1972 to have been used in over 10,000 classrooms and at over two dozen colleges and universities (Gillmore, 1972). During the past few years, however, students ratings forms and their associated support systems have become more widely available from institutions that actively, rather than passively, solicit interest in these forms and services. This more active provision of student ratings systems is exemplified by the support of John Centra's development, validation and distribution of The Student Instructional Report (SIR) by one of the nation's most prestigious evaluation institutions, the Educational Testing Service (Centra, 1972). Distribution of SIR has been sufficiently extensive to permit Centra and his associates (Centra & Creech, 1976) to sample ratings data from a pool of over 16,000 college classrooms. Another national distribution of a single ratings instrument involves the sponsorship by the W.K. Kellog Foundation of the INSTRUCTIONAL DEVELOPMENT AND EFFECTIVE ASSESSMENT SYSTEM (IDEA) developed by Donald Hoyt and his associates at the Center for Evaluation and Development in Higher Education of Kansas State University (Hoyt, 1969, 1975). This IDEA system was intended to be used by a national network of participating colleges and was used by approximately 400,000 students who evaluated 17,234 courses at 98 institutions since 1975 (Cashin & Slawson, 1977a, 1977b, p. 6). A third system, The Purdue Research Foundation's Cafeteria System (Derry, Seibert, Starry, VanHorn, & Wright, 1977; Starry, Derry, & Wright, 1973) has been adopted by at least 70 institutions and is providing service and licensing rights to anyone interested.

In Summation

Since 1971, there has been an apparent increase in interest and research dealing with student ratings in classroom instruction. Some of this upsurge seems related to an increased pressure upon administrators to employ measures of teaching effectiveness which are more open to empirical validation, while some of this activity is related to a recent and rapid increase in faculty development programs. A qualitative shift in the creation, use, and availability of rating systems has involved the active dissemination of a few, well-supported rating systems to large numbers of colleges and universities, and the consequent availability of large bodies of data which transcend the boundaries of individual institutions.

Student Ratings: Major Functions

Student ratings forms are sets of questions which ask students to describe or evaluate aspects of their college teachers and classes. These multiple-choice and/or open-ended questions usually request student opinion concerning teacher behavior, class procedures, course requirements, and student estimates of their own learning or interest in the class. The information produced from these student judgments, descriptions or opinions can serve many purposes, but the most commonly purported functions involve using the information to make decisions. Gillmore (1972) has labeled and described the three major decision-aiding functions as: (1) the normative (or summative) function involves using ratings data to compare faculty, either against one another or against established standards, so as to make or support personnel decisions concerning reappointment, promotion, tenure, salary increase, and teaching assignment, and thereby, also provide incentive to encourage faculty to do their best teaching; (2) the diagnostic (or formative) function involves providing faculty with information which can be used by them to improve their own teaching effectiveness; and (3) the informative function involves providing students with information which might assist them to choose classes and instructors. Smock and Crooks (1973) described ratings scales which were designed to produce three levels of information: (a) Level I questions are intended to produce comparable, global, concise, summary descriptions of teachers or courses; (b) Level II questions are intended to elicit descriptions of specific attributes of instructors or courses; while (c) Level III questions are usually open-ended questions designed to produce diagnostic information and suggestions for instructional improvement. Smock and Crooks also provide a form that included questions intended to elicit these three levels of information. The form presented as Appendix B in the sixth edition of McKeachie's Teaching Tips (1969, pp. 247-252) substitutes sets of polar adjectives and a rating line for the multiple-choice cues of the Smock-Crooks style of Level I and Level II items. Both forms provide an opportunity for students to provide diagnostic commentary (Level III) following each question. The great majority of other ratings forms apparently were designed to be electrically scored and, consequently, only provide opportunity for Level III general commentary on separate pages or on the back of the optically scanned answer sheet. Although two of the three Smock-Crooks levels of data generality do correspond to some degree with the normative (global, concise, summary, Level I) and diagnostic (open-ended, Level III) functions, the Level II type of items tend to make up the bulk of most ratings scales and are less clearly related to a single specific function. Grouped into scales, these Level II items provide normative data for personnel decision makers, and individually these Level II items may provide descriptive information for student use or diagnostic feedback for use by the faculty.

McKeachie (1969, p. 219) has identified a fourth and a fifth function for student ratings: "to improve student morale and stimulate student thinking about educational objectives". Ratings procedures also may sensitize faculty to various aspects of the educational process and thereby stimulate faculty thinking about educational objectives (function number six). Costin, et. al., (1971, p. 512) provided a seventh function:

course and program evaluation. Since ratings data describe aspects of educational process and outcomes, an eighth function could involve increasing knowledge concerning education, i.e. educational research.

Since student ratings forms and the resultant data are created and distributed by particular individuals and groups, these procedures could serve a number of idiographic functions for those involved with them. Some of the more likely functions include: the provision of visible evidence of research, administrative or service activity for the rating system operators; the transfer of individual responsibility for personnel decisions to an other-than-self agency; and/or the placation of the demands of accountability-oriented pressure groups. Data from Clark and Guba's (1976, VII, pp. 21-23) survey of college administrators caused these researchers to conclude that "the culture of institutions of higher education is predominantly idiographic (i.e., emphasizing the self-actualization of the professor) rather than nomothetic (i.e., emphasizing the goals of the institution)", and also that the predominant, collegial mode of governance "combines with the inexperience and lack of training of many administrators...to lead to decision-making processes...which do not meet the conditions of more classical models of decision making". Consequently, it might be useful to consider that the more frequently presented, nomothetic functions of student ratings also may be accompanied by other, less frequently described idiographic functions and associated motives.

The Reliability of Student Ratings

In their 1971 review, Costin, et. al., concluded that "students can rate classroom instruction with a reasonable degree of reliability" (p. 513). This conclusion was based partly upon: (1) internal consistency coefficients for full scales of .88 (Lovell & Haner, 1955), .93 and .85 (Spencer, 1965; Spencer & Aleamoni, 1968); (2) correlations of .69 (Cohen & Humphreys, 1960) and .93 (Harvey & Barker, 1970) between global item scores and total scale scores; and (3) half-term stabilities of .48 to .87 (Costin, 1968a) for subscales and a two-week stability coefficient of .89 (Lovell & Haner, 1955) for a full scale. Since 1971, the majority of subsequent data has supported the conclusions: (1) that almost all scales produced high internal consistency coefficients; (2) that almost but the "odds and ends" subscales produced reasonably high internal consistencies; (3) that the test-retest stability of scales, most subscales, most global items, and many of the more specific items are moderate to high; and (4) that most scale scores correlate well with the scales' global items. McKeachie stated this position in a letter (personal communication, 1 October. 1974) when he wrote that, "I know of no form that has a reliability of below .8, and most are well above .9 with classes of size 15 or more." Doyle (1975) reviewed a large number of reliability studies and also concluded that "the degree of freedom from random error indicated by these studies compares quite favorably with the reliability of well-constructed, commercially published achievement tests" (p. 35); that "student ratings are...or...can be made to be... very reliable in the sense of internal consistency" (p. 37); and that "those studies that have appeared...indicate a reasonable degree of stability for student ratings" (p. 37). Since 1971, an analysis of variance approach to reliability estimation (Centra, 1972; Winer, 1962) has employed comparison of between-teacher variance with within-teacher variance in order to produce reliability estimates which describe scales, subscales or single items. Intra-class correlation (Gillmore, 1973) has also been used to estimate item or scale reliabilities. While such intra-class correlations for a global item (Gillmore, 1973, p. 4) were computed to be only .222 for one sample of 200 classes and .246 for a similar sample when only one student was included in the class, for classes of nine students, the estimated item reliabilities surpassed .70. For classes of 14 students, reliabilities exceeded .80, and for classes of 31 or more, this item reliability estimate exceeded .90.

A point which is easily overlooked is that different uses of ratings data require freedom from different types of error, and consequently require different approaches to the estimation of reliability. If the judgments or opinions of individual students are one of the foci of a study in which these judgments are to be related to other student characteristics, the internal consistency of a ratings scale or subscale would be of crucial importance, and a .22 single item, single student estimate of reliability would be so low as to be useless. Enough additional items would have to be added to the scale or subscale to produce a single student rating which was likely to be more representative of the true opinion of the individual student. However, when one wishes to compare the opinions of an entire class with the opinions of other classes, the analysis of variance approach to reliability estimation may be more appropriate -- especially when there is reason to believe that the classes differ in their opinions or judgments —as would be expected when similar groups of students judge teachers who differ in clarity of presentation, fairness of testing procedures, or general effectiveness. If one were interested in the study of a characteristic presenting high intra-class variance, such as the "difficulty of the course" in a course catering both to majors and non-majors, an ANOVA or intra-class approach to reliability estimation might provide deceptively low results, and test-retest stability for individuals or relevant subgroups would be more useful as evidence that the teacher was unable to resolve a common instructional problem than a useful indicator of measurement error. The common finding of high correlations between total scale scores and responses to one or more global questions might support the use of the global items as summary indicators of overall teaching effectiveness, but this evidence hardly creates confidence in the use of individual items or subscores for diagnostic or explanatory purposes (fortunately, however. errors in diagnosis usually are much less expensive than are errors involving normative, personnel decisions).

Stability and Generalizability

If student ratings accurately reflect relatively invariant characteristics of teachers (general teaching effectiveness, enthusiasm, clarity of exposition, concern for students) or of courses (nature and value of course content), it would be expected that test-retest stability for items or subscales would be moderately high since the same group of evaluators would be providing opinions concerning the same teacher and courses. As reported by Doyle (1975, p. 39) test-retest stabilities (same section, course, teacher, term, scale, -- but different day) for most subscales range from .4 to .9, and full scale test-retest reliabilities often reach the high .80s and low .90s. It was unreported that many of the lower subscale stabilities included by Doyle and others are produced by the "odds and ends" subscales which tend to be collections of questions that are derived from scale factors which do not account for much of the total scale variance.

If most student ratings scales are intended to be reflectors of the relatively invariant characteristics of "teaching effectiveness", more demanding tests of stability would involve: (1) asking the same student observers, from the same section of the same course, to rate the same teacher, but using different rating scales; (2) asking different students to rate the same teacher on the same scale in the same course (by involving a parallel section of the same course); (3) asking different students to use the same scale to rate the same teacher in the same term but for a different course; (4) asking different students in different terms to rate the same teacher teaching the same course; and (5) asking different students from different terms to rate the same teacher on the same scale but without specifying courses. In effect, the various potential sources of ratings variance would be identified and attributed to teacher, course, or other potential causes. Some of these potential sources of variance would be held constant while others would be permitted to vary. Ideally, each facet (possible determiner of variance) or combination of facets would be manipulated throughout it full range in order to determine how much of ratings variance can be associated with the teacher (or even, with specific characteristics of the teachers' behavior, style or nature) and how much variance is attributable to various non-teacher facets of the college classroom. In reality, to date, various pools of ratings data have been searched to identify sets of student ratings which illustrate one or more of these possible tests for stability.

Stability Across Different Ratings Forms and/or Across Different Students

Three separate ratings instruments were used by 108 of 125 faculty at a typical liberal arts college (Sherman & Blackburn, 1975, p. 2). One form was provided by the college's Office of Instructional Research and was oriented toward the presentation of course content and proficiency in facilitating classroom interaction; the second form was provided by a student organization and focused upon the teacher's classroom organization and course requirements; while the third form was a semantic differential scale containing bipolar adjective-pairs relating to teacher attributes and personality traits. Since some students completed all three forms while some students completed only one of the three forms, different populations of students, under different conditions, and at different points over an 18-month time interval, rated the same teachers on different forms to produce three correlation coefficients of .77 (45 teachers) for the Research Office-semantic differential forms, .73 (63 teachers) for the Research Office-student produced ratings forms, and .86 (75 teachers) for the semantic differential-student produced forms. Aleamoni & Yimer (1973) investigated the relationships between ratings and other factors by: (1) asking faculty of the University of Illinois/Urbana-Champaign to nominate three good teachers; and (2) proceeding to collect available data on the 477 faculty that were nominated as effective teachers. In the process, the authors identified 40 nominated teachers who had been rated by students using the student "ADVISOR" during the 1969-1970 year who also had been rated through use of the Course Evaluation Questionnaire (CEQ) some time between 1966 and 1970. In spite of the truncated nature of the sample (only faculty nominated as good teachers were included) and the fact that data from a variety of each teachers' classes, taught over a four-year period, were averaged, the correlation between the whole scale CEQ and the instructor subscale of the ADVISOR was .58, while the course subscale of the ADVISOR and the CEQ produced a correlation coefficient of .52.

In these two studies, it was possible for data from some of the same students to be included in more than one of the mean ratings scores for any one teacher. This inclusion of the same students was precluded when Gillmore (1972) identified 100 instructors who each taught two sections of the same course during the Fall of 1970 and compared the item means (on the CEQ) produced by the two sections for each teacher. The resulting, intra-teacher item correlations ranged from .53 to .82, with a mean of .67. A similar study of faculty who taught two sections of the same class (Derry, 1977), used one of the global items of the Purdue Cafeteria Scale, "overall satisfaction with the instructor", and obtained convergence coefficients of .85 for 50 senior faculty and .63 for 55 teaching assistants. However, examination of the scatter plots and of the correlation coefficients produced by the middle 50% of these two groups of Purdue teachers, led Derry to state that "the use of item data (e.g., core item #5) to rank 'average' teachers (those in the middle 50% on University norms) leads to tenuous conclusions" (p. 15). Another approach which insures that independent sets of student raters are involved was accomplished by Drucker & Remmers (1951) who found correlations ranging from .40 to .68 between ratings item scores produced by current students and item scores produced for the same 17 teachers by former students who had graduated 10 or 15 years before. Centra (1973) produced correlations from .75 to .84 when ratings by current students were compared with global ratings of 23 teachers by recent alumni who had graduated within the past five years. These alumni produced correlation coefficients have considerable potential instability due to the small numbers of teachers involved.

Slightly larger numbers of teachers were studied by Marsh (1977) and reported by Hildebrand, Wilson & Dienst (1971), but both sets of studies examined only those teachers who were multiply nominated as "best" or "worst". Marsh asked graduating seniors to nominate "the most outstanding" and "least outstanding" teachers within their departments. Thirty-one multiply-nominated "most" and 31 multiply-nominated "least" teachers had their subsequent terms' student ratings examined and compared. Although the two groups of teachers differed most on the "overall instructor", global, summary item, next most on the global "overall" course item and on the "enthusiasm/interestingness" subscale, and did not differ on the difficulty/work load/pace subscale, the canonical correlation coefficient of the discriminate function analysis was .82. Hildebrand, et. al., (1971) reported that there was a high degree of agreement (chi square of 29.1, $p < .0005$) between the outcomes of two surveys of students asked to nominate their "best" and "worst" teachers even though the 57 teachers who were multiply nominated in both surveys were nominated in different years (1963-1966 for one survey and

1967 for the other). Subsequent global ratings items were reported by the students of 15 of the second surveys "best", 18 "worst", and 18 not nominated teachers. The results indicated that each group of teachers produced mean, global item scores that were significantly different than those produced by the teachers in the other two groups.

Evidently, different groups of students -- often using different ratings scales -- produce similar, but not identical, "effectiveness" ratings for the same sets of teachers.

Attribution of variance to teachers, courses and items. The multiple scale, alumni and nomination studies, although limited in number, do indicate that teachers and/or their courses elicit relatively stable indicators of student satisfaction, but the data analysis tend to confound the effects of two of the major possible sources of student satisfaction: (1) the general effectiveness of the teacher; and (2) the general attractiveness or perceived value of those courses that are most frequently taught by these teachers. Seiler, Weybright, and Stang (1977) analyzed the ratings obtained by 320 different instructors who taught 1,270 students during the autumn terms of 1972 and 1973. The instructors' ratings were selected for analysis if they taught two sections of the same course and a section of another course during one term, and if this pattern were similar during the following term. The highest, single "instructor rating" item convergence correlations were for the same course taught during the same year (r = .69), followed by the same course taught a year later (.63), a different course the same year (.54), and a different course a year later (.45). Because of the large number of instructors, these stability or convergence coefficients were quite stable themselves. A similar ordering of convergence coefficients occurred with all other items except those reflecting "instructors' availability outside the classroom," which were more similar within a particular term than within a particular course (as might be expected if faculty availability or unavailability were more dependent upon term than upon the course taught). Ratings of course characteristics tended to be less stable or convergent than ratings of instructor characteristics, and ratings of the instructors' verbal abilities appeared slightly more stable than other instructor characteristics (although the authors made no attempt to test the significance of these "differences" in stability).

Hogan (1973) used seven 5-item scales to collect data from two successive terms to produce comparable stability/convergence coefficients for the "global course and instructor scale" of: .67 for the same course and same instructor on successive terms; .40 for different courses with the same instructor on successive terms; and a median subscale correlation of .19 for different teachers of the same course. Bausel, Schwartz, and Purohit (1975) replicated and extended the Hogan approach (without referencing it) by analyzing data produced from three separate ratings forms that had no identical elements. As with Hogan, comparisons were made across successive academic terms. However, the data produced by any of the three forms was produced in different academic years and in different colleges of the same university. When all item correlations were averaged across the three scales, the resulting mean correlation coefficients were: .69 for the same course taught by the same instructor in adjacent terms; .33 for the same instructor teaching a different course; and .17 for the same course taught by different instructors. Comparable coefficients for global instructor items were not calculated by the authors, but appear to be, approximately: .78 for the same course and same instructor, but different terms; .36 for the same instructor teaching a different course; and .1 for the same course taught by different instructors. One of the three scales doesn't have a global course item, but the comparable course item coefficients for the other two scales were: .7 and .73 for the same instructors teaching the same course during different terms; .33 and .20 for the same instructor teaching a different course; and .18 and .21 for the same course taught by different instructors.

Generalizability theory. More recently, Kane, Crooks, and Gillmore (1976), Kane, Gillmore, and Crooks (1976), and Gillmore, Kane, and Naccarato (1977), have employed generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) to analyze the dependability of student ratings of instruction. The combined results of these three reports suggest that ratings' variance can be partitioned into approximately: 40% attributable to teacher effect; 6% attributable to course effect; and 54% attributable mostly to an interaction between teachers and courses, and slightly to a sections within course effect. Gillmore, Kane, and Naccarato (1977) produced their generalizability tables and variance estimates from the data collected using 11 general items from the University of Washington's Instructional Assessment System (Gillmore, 1974) with two sets of 42 pairs of classes representing the humanities, the social sciences and the physical sciences. Although the rating forms, the numbers of questions, and the class sizes of the Gillmore, Kane and Naccarato (1977), the Hogan (1973) and the Bausel, et al.

(1975), studies differ, an assumption of ten items and twenty students per class permitted Gillmore, et al., to use their generalizability equations to compare their results with those produced by earlier studies. The resulting generalizability coefficients (stability coefficients) in order of publication date (e.g. Hogan, Bausel, et al., Gillmore, et al.) were: (1) same teacher, same class -- .70, .69, .76 or .80 (.63); (2) same teacher teaching a different class -- .40, .33, .48, (.45); and (3) different teachers teaching the same class -- .19, .17, .05. The two correlations in parentheses were drawn from the Seiler, et al., (1977) study and appear constant with the other results.

The advantage of the generalizability approach to the estimation of ratings stability lies in its ability to permit estimation of stability coefficients in situations which are not immediately available. Gillmore, Kane, and Naccarato (1977), selected only two sets of data produced from eleven ratings items, 42 teachers who had taught two different classes within one academic year, and 42 other pairs of classes that had each been taught by two different teachers. Class size was equalized by choosing classes with 14 or more students and by randomly eliminating student data from classes with more than 14 students. From different combinations of the resulting components of variance, the authors were able to estimate stability coefficients for courses or for teachers for any number of items, students, courses, or teachers -- and even for all courses a teacher might teach or for all teachers who might teach a particular course. Several of the resulting conclusions were:

(1) Any particular course is unlikely to produce high enough stability coefficients to permit an evaluation of its effectiveness over all possible teachers, although adequate stability is possible if effectiveness is generalized over a particular set of teachers;

(2) If a teacher is likely to teach a variety of different courses and, consequently, a personnel committee wished to generalize effectiveness over all possible courses to be taught in the future, the decision to grant tenure should be based upon results from at least five different courses (not from five different sections of the same course) which each include at least five students (stability coefficients will be approximately .62 for five courses with five students per class and ten rating items; .71 if there are 20 students in each of the five courses; and still only .82 and .84 for 10 courses and 5 or 20 students per course, respectively -- from Gillmore, Kane, Naccarato, 1977, Table 3, p. 23);

(3) It is much more important to sample larger numbers of students than to sample larger numbers of items since increasing numbers of students increases stability of course-teacher evaluations much more than does increasing numbers of items;

(4) The large teacher-course interaction component seems to suggest that teachers should not be allowed to determine which of their courses be chosen to produce normative-summative data for personnel decisions; and

(5) This large interaction component also suggests that ratings of student satisfaction can be improved through more careful assignment of teachers to those courses in which they receive higher ratings.

Reliability in summary. The internal consistency of most ratings scales, and many subscales, often exceed .9, and, when lower, often can be increased by increasing the number of ratings items in a particular scale. However, this indicator of reliability produces a considerable overestimation of ratings stability when independent measures of teachers and/or courses are desired. Generalizability theory has begun to be employed to supplement and to extend the more traditional approaches to reliability. Conclusions from both traditional approaches and generalizability studies suggest that stability coefficients of .70 or greater are produced when the same teacher, teaching the same course, is rated by any fifteen or more students using virtually any four or more of the more general items found in the better rating scales. Course effectiveness measures for all possible teachers are unlikely to produce satisfactory stability unless very large numbers of teachers are evaluated; estimation of teacher effectiveness across all possible courses requires sampling from five to ten courses to achieve stabilities exceeding .7; while estimation of teacher effectiveness which reflects specific sets of courses can approach .85 under typical conditions. An apparently large course-teacher interaction suggests that many teachers elicit different rating scores when teaching different classes. There is also some evidence that experienced teachers produce greater stability of ratings, and that the rating stability is greater for teachers at the upper and lower quarters of the ratings distributions.

Validity of Student Ratings

Costin, et al. (1971) devoted approximately two-thirds (14 of 21 pages) of their review of student ratings to a consideration of the validity issues. Although this review dealt with a wide range of topics, much of it, and the core of similar reviews, represent an attempt to answer three questions.

(1) Do ratings items, subscales, factors and scales--and the procedures involved in their identification and creation--reflect what is known about relevant teacher/course characteristics, processes and products?

(2) Does the data produced when students respond to ratings items, subscales or scales correspond to data produced by other raters--or to data produced through use of other approaches to the measurement of the same or similar constructs?

(3) Are there particular characteristics of teachers (pleasing personality, grading policy), of students (major, expectations, interests, motivation, gender, class level, age, experience, intellectual ability or aptitude), or of courses (size, style, difficulty, academic discipline, course level, required versus elective, position in the major sequence, time of day offered), and/or ratings procedures (timing within the term, directions to students, question order, introductory comments by the instructor) which influence the data produced from the use of ratings but which are not believed to be actual or legitimate components of whatever the ratings are

purported to measure--in other words, are there situations in which ratings results are shifted upward or downward because of faculty manipulation, grading policy, preadministrative comments or directions to raters, or because of conditions beyond the control of many teachers (class size or class placement in the student's course sequence, pre-class student aptitude interest or ability, etc.)?

The first question reflects concerns about the content-related validity of student ratings; the second question includes concerns about criterion-related and construct-related validity; while the third question involves all three types of validity as they relate to the fair and unbiased use of student ratings in the normative or evaluative/ summative/ comparative functions.

Student Ratings Items, Factors and Scales: Construction and Content. Most faculty ratings scales include:

(1) Specific questions (Level II) which ask students to describe teacher attributes or behaviors, course characteristics, class procedures or processes, and/or learning outcomes; and/or

(2) Global items (Level I) which often ask students to evaluate the teacher, the course, or their own satisfaction or learning in comparison with previously experienced teachers, courses, or learning gains; and/or

(3) Open-ended questions (Level III) asking for student commentary or suggestions concerning the teacher or course experience; and/or

(4) Questions completed by the student to provide information such as gender, class level, major or minor, expected grade, G.P.A., reason for entering the class, initial interest in the class topics—or questions completed by the teacher to indicate class size, class level, department, goals, teacher workload, teacher's previous teaching experience, gender, or satisfaction with the ratings scale or items.

Some Scales (e.g. the Teacher Performance Inventory, Breiter & Menne, undated; and the Faculty and Course Evaluation Instrument, Sockloff & Deabler, 1971) contain no global items while other rating systems (e.g. use of the five summary descriptions of the Hildebrand, Wilson & Dienst short form, 1977, p. 18; or the front of the Student Description of Teaching, Wood, 1977) provide only global items, but the majority of ratings forms include both specific and global items, and many also make provisions for students to respond to open-ended questions and for faculty to add additional specific, global or open-ended questions. McKeachie's form (1969, pp. 247-252) and the Smock and Crooks (1973) physics department form are among the few well-known rating forms that intermix multiple-choice and open-ended questions in a format that precludes machine scoring of the multiple-choice items. Libbee (1976) has even described a ratings procedure entirely devoted to teacher-chosen, open-ended questions in which the analysis of the open-ended responses are accomplished by the class that produces the responses and then provided to the instructor and to the class within an hour of the time the questions were originally presented to the class.

Because of the great variety of different ratings scales, it is impossible to determine more than that the questions on any form represented at least the description of effective college education as viewed by the creator of the form. However, a considerable number of ratings forms represent definitions of effective teachers and/or courses that were agreed upon by a considerable number of people. Breiter and Menne (undated) described the steps involved in the construction of their Teacher Performance Inventory, and an edited version of these steps includes:

(1) The identification or creation of a group of items (92) believed to describe effective teaching;

(2) The ratings of these items for importance, usefulness and apparent validity by groups of students, faculty and/or administrators;

(3) The trial use of the more highly rated items with typical classes;

(4) The analysis of the student responses in terms of response distributions for items; the factor analysis or discriminate analysis to identify item groupings; the analysis of variance to estimate item discrimination coefficients; and the use of intra-class and intra-factor correlations to estimate item convergence across students and the internal consistency of any scales or subscales.

An additional effort to ensure content-related validity and administrative practicality often involves several scale revisions, each followed by data analysis and solicitation of student, faculty and committee suggestions. For example, the Student Instructional Report was built upon work on ratings forms at the University of Michigan in 1965, was first tested in expanded form in 1970, was shortened and revised for use in 1971, and was again modified for use with larger numbers of classes in 1972-1973 (Centra, 1972). The items of the Hildebrand, et al. (1971) scales were used with groups of teachers that had already been multiply nominated as "best" or "worst" teachers by two groups of students and by a group of faculty. Only those items that were effective in discriminating between best and worst teachers were included in the final form. The long and short versions of the Endeavor Instructional Rating Forms developed by Frey (1973a, 1973b, 1978) were modified and rewritten several times in order to achieve a combination of factor clarity and scale brevity. The widely used IDEA Survey Form (Hoyt & Cashin, 1977) is the result of three previous use-and-revision cycles: Form No. 1 was first used in 1969; and the current form is still under potential revision. The various IDEA revisions took place even though many of the original items were selected from scales that had had considerable use and analysis, and the first of the three trial forms was constructed with the advice of several faculty and faculty-student committees.

Although ratings forms are more easily improved while in their formative stages, and few ratings forms had prenatal periods as long as those available to the creators of SIR and the IDEA Survey Form, user dissatisfaction and/or discrepancies between the data produced by form use and patterns predicted by theory have led to the modification or replacement of ratings forms with relatively long histories of use. User complaints, poor sub-scale reliabilities and factor instability (Villano, Rosenstock, & Estes, 1974) were reported for the form used for over ten years at Pennsylvania State University, and a shorter, more factor stable form, the Instructor Evaluation Form (IEF) was produced as a replacement (Villano, 1977). An attempt to increase user satisfaction (face validity) while maintaining continuous possibilities for modification are represented in the Purdue Instructor and Course Appraisal or Cafeteria System (Derry, et al., 1977) which provides users with a catalogue of 200 items from which individuals, departments, colleges and/or universities may select those items that best match various models of effective teaching. A five-item core of three widely accepted, general, descriptive items and two global questions is available to provide users with a well-tested Level I anchor for normative purposes, while previously generated, item statistics are available for all items in a catalogue. User choice also is available with the IDEA Survey Form which not only provides spaces for up to twenty-five instructor-chosen questions (a common provision available with many scales), but also permits the teacher to indicate which of the ten possible, general course objectives are believed by that instructor to be "essential", which are believed to be important, and which are believed to meet neither criterion. Many of the remaining, specific items from the scale and the "overall evaluation" score are normed only as they relate to the instructor's chosen course objectives. This IDEA system also automatically compares item and subscale means that are produced by each class with means that are produced by a large pool of other classes (another common service), but also compares item and subscale mean scores produced by any particular class with those that were produced by classes which were similar in size (five size categories) and also similar in terms of five levels of student, self-reported pre-class desire to take the course (quite an uncommon option since few rating systems have stored enough class mean scores to permit comparison with large numbers of classes in each of the 25 cells of the 5 X 5 matrix). Somewhat more common is the provision of the score distributions or distribution characteristics (mean, median, standard deviation, centiles or deciles) achieved on key items or scales by faculty of different ranks or from different academic units. The availability of this data permits faculty or administrators to compare the scores of individual faculty with those of faculty believed to be in similar positions. The University of Illinois (Bussell, Costa, Spencer, & Aleamoni, 1971) has even published a set of computer programs for the processing of student ratings and similar measures, and the "Quest" utility program of this Mermac statistical package permits the inclusion of a single norm table that applies to all items and up to ten separate norm tables for each of a possible 20 multi-item subscores (pp. 68-85).

Do Ratings Scales Reflect Relevant Characteristics?

Sets of teachers' scores derived from student-ratings questionnaires could possibly be as reliable as is indicated by the results of the various reliability and generalizability studies (i.e. moderately reliable--as reliable as are many of the better achievement test scores) and still be invalid as indicators of effective teaching. Students may be responding to questions which permit a consistent or reliable response but which are unrelated to the truly crucial or central characteristics which differentiate good teaching from poor teaching. Even if descriptors of these crucial characteristics are included as part of a ratings scale, it is possible that responses to irrelevant questions will tend to distract from the effect upon the total ratings score that is produced by the responses to the "relevant" questions. The scale construction procedures that were described in the preceding pages have employed faculty and/or student consensus as a guide to the identification of "relevant" items. Even many of the scales that have been produced by a less demanding approach to consensual validation have often involved borrowing scales or items from the more carefully developed scales in a process that has become common enough to have had a name created for it--the "prestige-library approach". It seems appropriate to ask if there is a common model of the "effective college teacher" that underlies the "typical" rating form--and if so, does this common model appear valid?

Widlak, McDaniel and Feldhusen (1973) analyzed the results produced from 22 studies of different ratings forms and claimed that the descriptions of the college teachers presented by these forms could be organized into the three categories of: (1) the teacher as actor or communicator; (2) the teacher as interactor or developer of rapport; and (3) the teacher as administrator or director of learning activities. The authors' factor analysis of their own questionnaire (which had been designed to reflect these three dimensions or teacher roles) indicated that the majority of their form's specific, Level II items and both of their global, Level I items ("overall course" and "overall instructor") loaded into the actor-communicator factor which included items dealing with teacher organization and clarity as well as items reflecting teacher's ability to stimulate interest and motivate students. A rapport or interactor factor could be identified as could a text-test factor, but these factors accounted for less of the total score variance and were composed of items that correlated (.51 to .88) moderately to well with the two global items and with the other items of the communicator factor. Feldman (1976b) analyzed seventy-two separate studies involving different student ratings scales and concluded that student descriptions of the superior teacher across studies consistently included: stimulation of interest, clarity and understandableness of presentation, knowledge of subject matter, obvious preparation for and clear organization of the class time, and enthusiasm. Descriptors related to rapport, respect for students, friendliness and openness were also frequently cited by students as characteristics of the superior teacher although these characteristics were less closely correlated with the global assessments of particular outcomes or the global, Level I assessments of teachers than were the clarity-organization-stimulation characteristics. Feldman's (1976b) secondary analysis of nearly 50 factorial studies of teacher-ratings scales caused him to conclude that most descriptions of the superior college teacher clustered into three sets of items which parallel those described by Widlak, et. al. (1973): (1) the teacher as actor-communicator-presenter of information (overall evaluation; overall outcome of instruction; stimulation of interest; clarity and understandableness; preparation and organization of the course; and enthusiasm); (2) the teacher as facilitator-interactor-reciprocator (friendliness, respect and concern for students; openness to other's opinions, encouragement of discussion, intellectual challenge, availability, and helpfulness); and (3) a third, less focused, more dispersed cluster of items which tended to link the rapport facilitator items with the global and actor items and which can be characterized as teacher-as-Director/Regulator/Administrator (fairness of evaluation, difficulty of course, workload, classroom management, and clarity of course objectives and requirements). Feldman presented a two dimensional representation of his secondary analysis of the various factor analyses (Feldman, 1976b, Figure 1, p. 261) which resembled a map of the United States. The presenter-actor items, the global outcome items, and the other global items were centered around "stimulation of interest" (the Oregon-Nevada-Idaho border) in the northwest quadrant; the interactor-rapport items were distantly linked to the enthusiasm, clarity and preparation items of the stimulating-actor complex but were themselves grouped in the northeast quadrant; and the various administrator-manager items were loosely scattered between and south of these other two clusters.

An analysis of his own rating form by Frey (1978) has produced a similar pattern. Although the original Endeavor Instructional Rating Form (Frey, 1973a) was composed of 21 items which had been created to reflect seven separate factors, when Frey shortened his form by including only one item for each factor (Frey, 1978), the resulting scale factored into two global factors very similar to the Actor (advanced planning, presentation clarity, and self-reported gain in knowledge) and Rapport (class discussion was welcome, personal help was available) dimensions identified by Widlak et. al. (1973) and by Feldman (1976b). Although Frey (1978) combined his "had to work hard" item with his "pedagogical skill" (Actor) dimension and also combined his "accuracy of grading" item with his "rapport" (Facilitator) dimension, an inspection of his factor loadings (Frey, 1978, Table 2, p. 77) for these two items indicates that "grade" loaded only moderately and somewhat similarly on both major dimensions and that "work" loaded only moderately on the skill factor. In other words, Frey's data further supports the beliefs that: ratings items fall into two relatively clear clusters (Skill or Actor, and Rapport or Facilitator); that self-perceived learning gain clusters with the Actor dimension; and that a variety of items related to course difficulty or quality of examinations don't quite fit into either cluster.

Hildebrand, et. al. (1971) produced a five-factor ratings scale and also created a shorter form of their scale in which each of these five factors (scholarship, organization/clarity, rapport with group, rapport with individual students, and dynamism/enthusiam) was converted into a single, global question. Wood (1978) employed slight modifications of these five, global, factor-descriptive questions as the first five items of a "StudentDescription of Teaching" rating form which also included five other global items and thirty-five more specific biographic or Level II items. An unpublished factor analysis of results produced by using this form with nearly 2,000 classes and 40,000 students indicated that the organization/clarity item was the single item best correlated with the major factor that clearly appeared to reflect organization, clarity, clear objectives, preparation for class, course organization and "effective use of class time". A second factor included the two Hildebrand et. al. (1971) rapport items as well as the enthusiasm/dynamism and the motivation, interest, openness and "sense of humor" items, while two other, lesser factors reflected course difficulty and tests-grading-text-movies items. However, the most global items (the instructor, the total course, student accomplishment, want to take more from instructor - learn more about field) loaded on both the organization-clarity (skill) and rapport factors (with self-reported student accomplishment associating more heavily with the skill factor). This movement of motivational and enthusiasm items from the Actor dimension of Feldman's (1976b) analysis to the Interactor-Rapport dimension might possibly be explained by the facts: (1) that the majority of the responders to the Student Description of Teaching were female, upper-level undergraduates or graduates; (2) who were taking relatively small classes (95% contained fewer than 35 students); (3) in their major area (usually in the College of Education). This same general description of these students might also explain the shift of the global responses from Feldman's position within the Actor cluster to an intermediary position between the skill-clarity cluster and the rapport-motivator dimension.

Gaff (1973) reported that when college seniors were asked to nominate the teachers who "contributed most to their educational and/or personal development" (p. 221), the nominated teachers differed from their non-nominated colleagues not on the dimension of classroom clarity/organization but rather on a dimension more similar to the dimension of rapport or willingness to interact with students. Feldman (1976b) similarly reported that when students were asked to provide a description of their ideal or best teacher, rapport descriptors such as friendliness, helpfulness, and openness to other's opinions dominated the descriptions, but that these descriptors were less useful than the clarity-organization descriptors in the prediction of global assessments of actual teachers.

The picture that emerges from these analyses is of: (1) the ideal classroom teacher who is knowledgeable, clear, enthusiastic, well-organized and able to present information in an understandable manner--an enthusiastic and well-organized lecturer who may or may not be good at leading discussions and interacting with students; and (2) the personal guide and mentor who is willing to interact with students beyond the confines of the classroom and who may or may not be an organized lecturer. The Actor and Interactor/Facilitator designations appear to be particularly appropriate labels for these two versions of the student's "ideal teacher", and the consistency with which these two models arise from various ratings scales provide some evidence of the content-validity of those ratings scales that reflect these two models of teaching. Teachers who epitomized the "Actor" model would be expected to garner high global teacher evaluations in situations where classes were primarily a series of information-dispensing lectures, while those epitomizing the "Interactor" model would be expected to elicit high global teacher evaluations under conditions in which transmission of information was secondary to other considerations or in situations in which discussions replaced or substituted for the lecture. Hildebrand, et. al. (1971) reported that multiply-nominated "best" teachers who were rated especially high on student interaction "tended to be giving small lecture-laboratory classes, were particularly favored by female, upper division or graduate students ... who valued Contribution to General Development ..." (p. 33). The "best teachers who were rated high on Scale 2, Organization/Clarity, tended to be giving large lecture or lecture-laboratory classes, were particularly favored by male, lower-division students with high Upward Mobility/Security orientations who valued the Transmission of Fundamentals..." (pp. 33-34). "Best" teachers who were rated relatively low on the organization/clarity scale were particularly favored by female seniors who valued "self-knowledge/ humanism" and "contribution to general development".

If this dual model of the ideal teacher is valid, Feldman's (1976b) finding that the Actor-role items best predict global ratings of college teachers would imply that most college classes were oriented toward the lecture-transmission of information rather than toward the discussion-interactive elicitation of knowledge or facilitation of general development. In an analysis of the usefulness of classroom observation measures, Batchelder and Keane (1977) randomly selected a sample of 48 of 220 college teachers at an Eastern state college and discovered that teacher lecture occurred for over 83% of the observed class time (92% of the time for natural science teachers, 84% for social science teachers, and 74% for humanities teachers). A wide range of other descriptions of college teaching (see Cole, 1978 for a listing of descriptions of and guides to "effective college teaching") similarly support the position that much of the classroom instruction by most college teachers involves the presentation of information--usually via the lecture approach.

McKeachie and Kulik (1975, pp. 165-209) performed a secondary analysis of the studies that compared the lecture and the discussion methods of college teaching as to their effects in producing three types of course outcomes: (1) knowledge measured through use of a "factual examination"; (2) "retention and higher-level thinking"; and (3) "attitude and motivation". Through use of a comparison across studies, these authors produced evidence for the superiority of the lecture approach in producing factual knowledge but inferior to discussion in producing the other two outcomes. Gage's (1978) recent analysis of successful teaching (to maximize the achievement of third-grade students) clearly reflects a teacher who is able to manage the classroom activities of students so as to maximize the time that they spend upon a variety of "academic learning tasks" and minimize the amount of time devoted to listening to directions or working on activities that are too difficult to accomplish or too easy to learn from. This description of the "effective grade-school teacher" emphasizes the third dimension of the Feldman (1976b) and Widlak et. al. (1973) teacher model--the teacher as the Manager of classroom activities.

The McKeachie and Kulik (1975) and Gage (1978) analyses can be combined with the Feldman (1976b) analysis to produce a process-outcome model of effective teaching in which: the Actor dimensions of teaching would be most associated with outcomes related to the transmission of knowledge, the Interactor dimensions of teaching would be most associated with outcomes related to retention, higher-level cognitive abilities, and self-understanding; and the Classroom Management dimensions would be most associated with outcomes related to the student production of skilled acts (such as reading, writing, computation, and the production of graphic arts) which require a considerable degree of guided drill and practice. Consequently, according to this model, global ratings of teacher or course effectiveness, student self-reports of learning gain, and external indicators of learning are likely to be most closely associated with ratings of the Manager dimension only when both faculty and students commonly perceive the course outcomes to be related to skills which require practice (e.g. shop classes, some speech classes, studio classes, and other classes oriented toward the development of observable skills). College level, studio classes are relatively rare (only 37 of 708 classes surveyed by Hoyt and Cashin, 1977), and college-level classes oriented toward the transmission of knowledge are apparently very common. Only quite recently (Hildebrand et. al., 1971 and Hoyt and Cashin, 1977) have faculty and students involved in student ratings been asked to characterize their classes as to their objectives, and the Hildebrand, et. al. ratings scales do not reflect the Manager dimension to any clear degree while the Hoyt-Cashin form only lists two skill-like outcomes, the ambiguous "developing specific skills, competencies and points of view needed by professionals in the field most closely related to this course", and "developing skill in expressing myself orally or in writing". The empirical validation of this three-dimensional, outcome-process model of effective teaching would seem to require asking students and faculty to provide information concerning the nature of course goals--information not now provided by most ratings scales or procedures.

The convergence of ratings items into three sets of factors provides one indication of the content validity of scales and the major subscales as representations of a common model of the effective college teacher. The close association between the actor dimension of the model and the global ratings of most college teachers appears understandable if most college teachers employ the lecture method to meet the major goal of transmitting information. An increase in the relationship between global evaluations (or student estimation of learning gains) and the rapport dimension would be expected when the class outcomes were attitudinal rather than cognitive in nature, and such a shift is evident, to some degree, in the Hildebrand et. al. (1971) data. The original development of the IDEA system also involved the prediction of different types of class goals, and Hoyt and Cashin (1977) also found that their rapport-related, student involvement dimension was a better predictor of the goals of self-understanding and effective communication than of the more cognitive goals related to gain in factual knowledge, gain in knowledge of principles and theories, and gain in knowledge concerning applications. These three cognitive gains were best predicted by ratings items associated with clarity of communication. Although such evidence is limited, it does provide some support for the construct validity of student ratings of college instruction.

Criterion-Related Validity

Doyle (1975) has suggested an "incremental approach to conceptualizing validity for student ratings" and has stated that on the first level of meaning, "ratings are as valid as they are reliable, and in the absence of contradicting or clarifying information these observations, perceptions, and judgements can be taken as meaningful in themselves" (p. 1). A second level of validation is to be derived from subjective procedures which involve gathering opinions as to item or factor importance from students, faculty and other interested persons and also involve examining item covariation and the internal structure of ratings scales. Doyle's (1975) third level of validity (or meaning) is estimated through the use of a set of more objective procedures by which data produced by the use of items or rating scales is compared with other indicators of the constructs purportedly reflected by the ratings items or scales. Direct objective validation, for example, would entail the comparison of ratings of teacher knowledge with other indicators of teacher knowledge or the comparison of student self-reports of gains in learning with scores on tests that reflected such learning. An example of indirect objective validation would involve the demonstration that ratings of increased interest in a course or subject would be associated with a subsequently increased entry into similar courses or that higher ratings of course or teacher characteristics believed to be associated with increased learning would, in fact, be correlated with other indicators of increased learning. This empirical or criterion-related approach to validity has produced a considerable degree of controversy which was highlighted when Rodin and Rodin (1972) published a report in Science that was headed with the summary statement that "students rate most highly instructors from whom they learn least" (p. 1164). The eleven "instructors" were teaching assistants who met with the second author's undergraduate calculus students for two days each week to supplement the three hours of lecture presented each week by the course director, Dr. B. Rodin. The rating was a mean response to the question, "What grade would assign to his total teaching performance?" The measure of student learning was the actual number of a possible set of 40 "paradigm problems" that students were able to complete with no error. Each student was permitted to retake up to six variants of each problem with no penalty. The product-moment correlation between the assistant's mean global rating and the number of problems completed by his students was a -.754 and caused the Rodins to suggest "that students are less than perfect judges of teaching effectiveness if the latter is measured by how much they have learned" (p. 1166). The mean ratings of all but two of the eleven "instructors" exceeded "B" (3 points or higher) and most of the observed negative relationship between ratings and problem completion involved the two outriders--the single low-rated instructor (mean rating of approximately 2.2 or C) whose students completed the most problems and a very highly rated assistant (apparent rated as A by all of his students) whose students completed the fewest number of problems. It was not very clear why all (or most) students, regardless of the actions of their course assistant, failed to complete one of each of the 40 types of "paradigm problems" that defined the major objectives of the course. Doyle (1975), Frey (1973b) and Gessner (1973) have all identified other weaknesses of the Rodin and Rodin (1972) study.

Glass (1976) claimed that "a study with a half-dozen design and analysis flaws may still be valid" (p. 4), and Gage (1978b) has suggested that all "educational research is likely to be fallible, flawed, and open to criticism of one kind or another ... thus the path to increasing certainty becomes not the single excellent study that is nonetheless weak in one or more respects, but the convergence of findings from many studies that are also weak but in many different ways" (p. 233). Previous to the Rodin and Rodin (1972) study the most commonly reported relationships between ratings and test scores were positive but only moderate at best. Remmers, Martin and Elliot (1949) reported a correlation of .27 between ratings of graduate assistants and the average grade attained by the assistant's students. Elliot (1950) produced a similar relationship (r = .24) a year later. Remmers, Martin and Elliot (1949) and Rodin (1972) rated graduate assistants who were responsible for only parts of a multisection course, employed pre-measures of ability to permit an adjustment for initial differences in student ability, involved students who may have self-selected their sections and section-leaders, and only attended to results from those ratings items that appeared most valid while not attending to data derived from items that were equally available. McKeachie, Lin and Mann (1971) employed a more sophisticated theory to test the hypotheses that ratings scales reflecting structure-clarity would produce results that correlated with test scores reflecting knowledge while scales reflecting interaction and feedback to students would predict scores of a test designed to measure higher levels of cognitive abilities. Male and female students were also expected to produce somewhat different outcomes with females more aided by instructors rated high on warmth and rapport and males more helped by instructors rated high on group interaction. As with the previously mentioned studies, the set of five studies reported by McKeachie et. al. (1971) focused upon instructors who taught different sections of multi-section courses, but unlike the previously mentioned studies, these instructors apparently had full responsibility for the presentation of the course material. In two of the five studies, the instructors were clearly labeled as graduate students, but the instructors were not identified in the other three studies but were probably graduate students for at least two, and perhaps for all three. Since McKeachie, et. al. (1971) reported the relationships resulting from the comparison of either four or six ratings subscales with from one to three other criteria in each of five studies (all of which reported results from males and females separately), 118 separate correlations between ratings and test scores were reported for the five studies. Approximately 36% of these correlations were negative, and 64% were positive. Of the thirteen correlations that were reported to have reached significance at the .05 level (simple "t" tests), twelve were positive and one was negative. McKeachie and his associates concluded that the results, taken together with the results of earlier studies, "did not invalidate the use of student ratings as one source of evidence about teacher effectiveness, but are less convincing than we had hoped for" (p. 444). There was only one example of an attempt at direct empirical validation of student ratings embedded in the five McKeachie, et. al. (1971) studies. In Study 5 (pp. 442-443), one of the criterion measures was a 14 item attitude sophistication questionnaire (with unreported reliability) which was believed to differentiate between the unsophisticated beliefs of the naive layperson and the sophisticated beliefs of the economist. The relevant ratings scale was believed to reflect the student's self-report concerning their change in beliefs (presumably toward those held by a sophisticated economist). The mean class scores derived from the student rating scale and from the attitude questionnaire correlated positively and significantly for the female students ($\underline{r}$ = .44, $\underline{p}$ < .05), but only trivially for the male students in the 18 economics classes ($\underline{r}$ = .09). Every other of the remaining 116 reported relationships would have to be classified as results of attempts at indirect empirical validation since each involved the assumption of a linear relationship between some student-reported characteristic of a class or teacher and the scores produced by the use of one or more tests of cognitive gain.

One of the major problems with the multi-teacher, multi-section validity study lies in the relatively small number of sections into which a course can be divided and the even smaller numbers of instructors teaching these class sections. The McKeachie et. al. (1971) studies averaged fewer than 15 instructors per study, Remmers et. al. (1949) and Rodin and Rodin (1972) each reported data derived from eleven graduate-assistants to the course director, Elliot (1950) was fortunate in collecting results from 36 chemistry instructors, but Mason and Mason (undated) were only able to report results from two instructors for each of two common courses. Morsh, Burgess and Smith (1956) reported upon a study that was unique in several ways: (1) they obtained data from the students of 106 different instructors who all taught the same eight-day course, in exactly the same physical surroundings and using the same teaching facilities; (2) each instructor taught two separate classes of the same course--the second section a month after the first; (3) objective measures of instructor knowledge of subject and general intelligence were available for all instructors; and (4) supervisor and peer ratings also were gathered for each instructor. No other widely reported study of student ratings, before or since, has involved so many instructors teaching the same course, and no other study involving adult students has gathered both ratings and achievement data for successive sections taught by the same instructors. Student global ratings of their instructor were correlated with combined achievement gains (r = .40), and student ratings of their instructors' teaching ability also were correlated with achievement gains (r = .46). Student ratings of the instructor's knowledge of the subject were correlated with one of the two student achievement scores (.22) and not with the written measure of student achievement, but were related to the instructor's score on an objective measure of their knowledge of the course content (.43). Student ratings of the instructor as a friend were not correlated with achievement gains, although ratings of the instructors' "understanding of students" was correlated with student learning to some degree (.28). All correlations represented product-moment comparisons between the mean achievement scores achieved by an instructor's two separate sections and the mean of the sum of two different achievement scores (adjusted for initial student ability) also achieved by each teacher's two different classes. Although the various supervisor and peer ratings of the instructors were moderately intercorrelated (.52 to .77), none of these judgments of instructor effectiveness were correlated with student achievement. Student ratings of an instructor by one class correlated moderately (.41 to .51) with those of the instructor's other class, as did the achievement scores produced by the two classes (.55). Because of the large number of instructors (106), all correlations exceeding .19 were likely to have been significantly different from chance (at the .05 level). The overall pattern of results indicated that the student ratings produced by each instructor's two sections exhibited somewhat less convergence than would be expected from college teachers, but a degree of convergence similar to that produced by graduate-teaching assistants (Derry, 1977), and that the relationship between the overall ratings (or the ratings of teaching ability) with the more objective measures of course outcomes were moderate and as high as could be expected considering the only moderate convergences of the measures across the two sets of different classes. Since Morsh, et. al. (1956) did not provide data concerning the amount of experience their instructors had had with teaching the hydraulics unit, it is not possible to attribute the only moderate ratings or learning convergence to lack of instructor experience or stability or to other factors. It also is difficult to determine to what degree the results of an eight-day course in aircraft mechanics, taught in an Air Force Technical School, can be generalized to the "typical" college classroom.

Shortly after the Rodins (1972) had reported a negative relationship between student ratings and student accomplishment, Frey (1973) published a critique of the Rodins' position and also reported a set of positive correlations between final examination scores (adjusted by regression on aptitude scores) and student ratings in two, multi-section calculus courses which were taught by regular faculty rather than by graduate students. The correlations between mean section ratings of student accomplishment and mean (regressed), section scores on common final examinations were .84 for the course taught by eight regular faculty and .90 for the course taught by five regular faculty. Correlations between teacher presentation and exam scores were .91 and .60, and were .65 and .73 for grading fairness and exam scores, .87 and .36 for course organization and exam scores, and only .14 and .49 when teacher-accessibility was compared with exam scores. Although the student ratings were, atypically, collected by mail after the students had completed the course, Frey later produced similar results when he surveyed another multi-section calculus course two years later (1976). In this later survey, half of the students in each of the seven sections were surveyed during the last week of the term while the other half were surveyed during the first week of the subsequent term. An analysis of variance provided evidence that the results of the two surveys were not different. Frey, Leonard and Beatty (1975) employed a common ratings form (that was similar to the one previously used by Frey) to compare the ratings results and final examination scores associated with the regular faculty teaching three multi-section courses at three different universities. Instead of mailing the ratings forms to the students, as in the other Frey studies, the forms were distributed in class during the last two weeks of the term. The correlations between ratings factors and the final examination scores for each faculty member were somewhat lower than in the other studies by Frey (1973, 1976), but the student accomplishment scale correlated positively and moderately (.50, .64, .61) with the exam scores in the three different universities, while the organization-planning factor (.85, .70, -.02) and the presentation-clarity factor (.81, .74, .18) also tended to correlate with final exam scores. Frey (1978) reanalyzed the data derived from the rating scales used in his previous research and concluded that his ratings scales appeared to reflect the two major global factors of "pedagogical skill" (student self-reported accomplishment, presentation clarity and course planning) and "rapport" (class discussion, fair grading, teacher availability). By analyzing his previously gathered data (Frey, 1973, 1976; Frey, Leonard & Beatty, 1975) in terms of these two factors, Frey (1978, pp. 86-87) concluded that the median correlation between an instructor-elicited rating of "pedagogical skill" and the exam scores of his or her students was .81 while a similar median correlation between "rapport" and exam scores was .29.

Another pair of studies produced results that were somewhat similar to those produced by the Frey series. Sullivan and Skanes (1974) compared the global overall ratings of instructors teaching 10 first year, multi-section courses with the mean final examination scores of each teacher's students. The median correlation was .41, and the mean correlation was .39. One of the courses had 40 separate sections that were taught by 27 full-time instructors and 13 part-time, graduate assistants. Although the correlation between ratings and exam scores for the 40 instructors of this course was .40, the correlation increased to .53 when only the 27 full-time instructors were examined, and to .69 when the first-year, full-time instructors were also eliminated from the analysis. Nearly three and a half years later, Centra (1976) investigated seven, similar, first-year, multi-section classes at the same university--Memorial University, Newfoundland. Centra and Rose used the Student Instructional Report (Centra, 1972) to gather student opinions from the 7 classes with their 72 sections taught by 44 full-time, regular faculty. The mean and median correlations (between mean final examination scores and mean ratings for each section) were, respectively: .45 and .60 for the overall teaching effectiveness, global rating item; .42 and .61 for the "value of the course", global item; .38 and .49 for the course organization subscale; .39 and .47 for the "general quality of lectures", global item; but only .22 and .30 for the rapport-interaction subscale. For five of the seven multi-section classes, the results derived from the global "teaching effectiveness", "lecture", and "course value" items correlated moderately to strongly (.4 to .9) with the mean exam scores for the sections, while the correlations were slight for the sixth course (.23 and .30) and slightly negative (-.16 to -.19) for the seventh course. The course-organization subscale tended to produce slightly lower correlations with exam scores while ratings of rapport, quality of examinations, course difficulty and quality of reading assignments were weakly and less consistently correlated with exam scores. An analysis of the relationship between mean student achievement and ratings of overall teacher effectiveness for the class which was subdivided into the largest number of sections (N = 44) did identify one particular teacher that, while one of the most highly rated teachers, produced student exam scores that were lower than those of all but two, much lower-rated colleagues.

The Centra (1976) study was, in many ways, a replication of the Sullivan and Skanes (1974) study and produced similar results. In 1978, Costin reported a 1973 comparison of the mean class ratings and mean final examination scores of 25 experienced (2nd term) graduate assistants who were responsible for the classroom presentation of a psychology course that was structured by others. Costin replicated his own initial experiment in the three following years (1974 to 1976). The rating scale consisted of five items: a global "all-around ability" item; a stimulation item; an "interesting presentation" item; an "explained clearly" item; and one item concerning teacher's skill in observing student reactions. The resulting correlations between the teacher's total ratings and the exam scores of his or her students were .52 in 1973, .56 in 1974, .46 in 1975 and .41 in 1976. There were 21 teachers in 1975 and 25 teachers every other year, all of these correlations were significant at the .05 level, two were significant at the .01 level, and all of the graduate assistants had taught the class at least once before the examined term.

Marsh, Fleiner and Thomas (1975) reported a similar study involving a multisection, introductory course in computer programming. However, Marsh, et. al. (1975) combined the results from two successive quarters instead of using the second quarter to replicate the investigation performed in the first quarter. The resulting study involved 18 different sections of the same course, taught by an unknown number of instructors who were of unknown experience but were categorized as "instructors, usually graduate students" (p. 835). The objective criterion measure was a common, "standardized", final examination of unknown characteristics, and the ratings form contained 46 items that factored into seven scales:

(1) a combination accomplishment and professional orientation scale (.30); (2) a presentation-clarity scale (.43); (3) a workload/difficult scale (-.02); (4) an instructor enthusiasm-rapport scale (.18); (5) an instructor approachability scale (.08); (6) a course organization scale (.55); and (7) a value of readings scale (.21). The decimal fractions in percentages were the correlations between the section means on the rating scales and on the final examination. Section means for three global items were also correlated with mean examination scores: (1) overall instructor (.44); (2) overall course (.30); and (3) a second "overall instructor evaluation" (.42). The optimal multiple correlation between ratings items and section exam scores was .74. Unfortunately, as was true in almost all similar studies, there were no indicators of the correlations existing between scales or items and no report of reliabilities for any of the measures.

An unusual study was reported by Gessner (1973). Instead of the usual approach in which different students were enrolled in the same multiple-section course and different teachers taught different sections, in Gessner's investigation, ten different medical school faculty each taught one or more of a course's 23 separate subject areas to the same 119 students. Consequently the effects of course content and teacher were completely confounded. Each of the 23 subject area/teacher combinations were rated as to content, organization and presentation; and these ratings were compared with two external indicators: (1) scores on those questions of the departmental examination that were believed to reflect the content area; and (2) the scores on the questions of the National Medical Board Examination that were believed to reflect the various content areas. The latter criterion was modified to reflect the local students' superiority to the national sample in order to attempt to equalize the difficulty level of sets of questions reflecting each area of content. No such attempt to equalize the difficult level of the departmental examination was made and, consequently, differences in student success in answering different segments of the departmental examination were as attributable to different levels of question difficulty as to different levels of student knowledge. The correlations between student ratings of the organization and presentation of a content area and the comparative success of students on the related segments of the National Medical Board Examination (given 5 weeks after the end of the course) ranged from .62 to .74, depending upon the ratings item examined. These correlations were similar in magnitude to the correlation of .68 between total student scores on the two criterion measures--the Board Examination and the departmental examination. However, when the items of the departmental examination were subcategorized according to the various content areas, the correlations between ratings and subsections of the departmental examination were slight (.11 to .17) as were the correlations between analogous subsections of the departmental and Board examinations (r = .18). Gessner (1973) concluded that "both student ratings and class performance on national normative examinations are valid measures of teaching effectiveness" (p. 569), and he discounted the low relationships between ratings and subsections of the departmental examination because of the confounding of item difficulty and the effectiveness of the teaching of the content unit. Unfortunately, Gessner failed to report the reliabilities of the 20 separate subsections of either of the two examinations, and it is entirely possible that the reliabilities of such necessarily short measures were so low as to preclude any subsequent test for validity.

Ratings vs. Examination Scores: Conclusions

If one were to depend entirely upon the Rodin and Rodin (1972) report and upon the studies that preceeded it, one would have to conclude that, except for data from a study that took place in a military etting, the relationships existing between student ratings and course examinations were occassionally positive, occasionally negative, often mixed, and usually slight. In the contrast, the study by Morsh et. al. (1956) and the majority of the studies that postdated the Rodins' report tended to present a relatively consistent pattern of relationships in which the examination scores of a teacher's students correlated, at least moderately, with: (1) student self-reports of learning gain; (2) global student evaluations of teaching effectiveness or course value; and (3) specific student evaluations of teacher clarity, organization, pedagogic skill or similar descriptors of the teacher as lecturer. Gessner (1973) and Sullivan and Skanes (1974) offered one possible explanation for this discrepancy in outcomes when they suggested that the "teachers" of several of the earlier studies (e.g. Elliot, 1950; Remmer, et. al. 1949; and Rodin & Rodin, 1972) functioned as course assistants to a clearly visible master teacher rather than as apparently independently-functioning teachers. Sullivan and Skanes (1974) also suggested that even independently functioning teachers who were inexperienced or working part-time would be less likely to focus upon student achievement or produce ratings that were associated with achievement. Sullivan and Skanes tested this hypothesis and produced evidence that the relationship between effectiveness ratings and test scores of a set of experienced, regular faculty was moderate and positive while the relationship produced by inexperienced or part-time faculty was nonexistent. This demonstration has not been replicated since previous investigators failed to indicate the experience level of their "teachers" and subsequent experimenters have tended to focus upon more experienced teachers. For example, Costin (1978) studied the outcomes of the second semester of four successive years to insure that "each teacher would have had at least one semester to become familiar with the course" (p. 87), and Centra (1976) specifically reported that "none of the teachers was a graduate teaching assistant, and almost all were experienced teachers" (p. 4). Although not specifically suggested by Sullivan and Skanes (1974), it is also possible that the inexperienced teachers or graduate assistants employed in many of the earlier studies may have been in the process of developing a teaching style and therefore have been less stable in their ability to elicit either ratings or student achievement. Derry (1977) reported that 55 graduate assistants, teaching two sections of the same course, produced a convergence coefficient on a global rating item, of .63 while 50 senior faculty, teaching two different courses, produced a convergence coefficient of .85 on the same global ratings item. Wood (1978, Note 1) similarly reported that global ratings of experienced faculty teaching different courses ranged from .71 to .81 across successive years, while those of graduate students, teaching the same course across adjacent terms, ranged from .43 to .52.

Wood (Note 2) attempted to examine some of the relationships between ratings and achievement elicited by graduate student-assistants by gathering student ratings and test scores from a multi-section, freshman-sophomore, college course that was jointly taught by one experienced master teacher and 43 graduate assistants. Although the master teacher functioned as the course director, wrote the midterm and final examinations, and presented one hour of lecture each week (the same, hour-long lecture was presented several times each week to accommodate the large number of students), the graduate assistants met with their sections for four hours each week to discuss and apply the course content to specific situations. The graduate assistants also directed, supervised and evaluated a series of five short speeches or similar performances, and the grades for these performances were combined with the scores on the midterm and final examinations to produce the course grade. The graduate assistants each met with their own groups of students for four of the five hours that the course met each week, had considerable autonomy of function, and actually presented classes more similar to seminar classes than to the laboratory or review sessions typical of several of the earlier studies.

A seventy-item student ratings form had been developed for use with this course, and it was completed by students during the last week of each of three successive academic quarters. Separate factor analyses were performed for each of the three administrations, and each produced six similar factors which could be labeled: (1) general value of the course; (2) effectiveness of the master teacher-lectures; (3) effectiveness of the graduate-assistant and seminars; (4) quality of the texts-readings; (5) workload; and (6) self-reported learning-accomplishment. Each of the six ratings scales and each of the midterm and final examinations produced alpha reliabilities that exceeded .80. The ratings subscale scores were created by adding together the scores of those items which loaded above .6 on a factor, and the resulting subscales were intercorrelated to various degrees. In spite of the fact that the graduate assistant had four times as much class contact with students as had the course director-lecturer, the ratings of course value, student accomplishment, texts-readings and workload were correlated nearly twice as strongly with the ratings of the master-teacher than with the ratings of the graduate-assistant for each of the three terms. In order to determine if the graduate assistants had an influence upon either the ratings or the test scores of their students, a series of one-way analyses of variance employed the graduate assistants as the dependent variable. There were no significant differences across graduate assistants, in any of the three academic quarters, for the midterm scores or final examinations scores. There also were no differences across graduate assistants for the premeasures of student cumulative grade point average or American College Test scores; there were slight but significant differences for the ratings of the course director, text-readings, value of the course and student accomplishment (eta^2 from .08 to .25); and there were stronger and highly significant differences across graduate assistants on the graduate assistant ratings subscale (eta^2 values of .43, .40 and .41) and on the total score derived from the five student speech-performances which were supervised and graded by the graduate assistants (eta^2 values of .66, .75 and .42). The convergence coefficients for the graduate assistant ratings across terms I and II was .18, while it was .56 across terms II and III and .63 across terms I and III. Similar convergence coefficients for other ratings subscales for the graduate assistants averaged .14 and ranged from -.28 to .58. Although the correlations between graduate assistant effectiveness ratings and the test scores of their students should be meaningless because of the lack of graduate student influence upon student test scores, the resulting correlations were .44 ($p < .04$), for term I, -.18 ($p < .23$) for term II, and -.26 ($p < .09$) for term III—and the corresponding relationships between the test scores and self-reported accomplishment for each of the graduate assistant's students were .44 ($p < .04$), -.53 ($p < .01$) and -.40 ($p < .02$). If an investigator had failed to discover that there was no relationship between assignment to a particular graduate assistant and student test scores, several of these apparent relationships would seem to justify judgments concerning the indirect (graduate assistant ratings vs. total test score) or direct (student selfreported accomplishment vs. total test score) empirical validity of student ratings.

Stability of Test Scores

Some of the lack of perfect correspondence between ratings of the effectiveness of experienced teachers (or student self-reports of their own accomplishment) and student test scores can be explained as a possible lack of correspondence between student goals and course goals, but some discrepancy can be attributed to a similar lack of correspondence between course or teacher goals and what a course examination actually measures. Extensive training in test construction is not a necessary part of the background of many college faculty, and even experts in evaluation find it difficult to construct reliable measures that reflect their most important course goals. Most of the validity studies focused upon the relationship between ratings and test scores have failed to report reliability estimates for their test scores, and several of those that have employed more than one measure of learning have not reported the relationships existing between the different measures. Gessner (1973) reported a correlation of .68 between two objective measures of learning, and this relationship was similar in magnitude to the relationship existing between various student ratings items and one of these measures (.62 to .74). Morsh, Burgess and Smith (1956) also employed two student performance measures in their study, a set of four seventy-five item, multiple choice, alternate forms of a test and a four-hour, job-sample performance test. They reported that the two measures "have about twenty-five percent common variance" (p. 88) which could, presumably, be translated to an intercorrelation of approximately .5--which is not too different from the reported .46 correlation between student global ratings of their instructors' teaching ability and student combined test scores. The study reported by Wood (Note 2) involved midterm and final multiple-choice examinations and five performance scores. The correlations between the midterm and final examinations were .64 (N = 834), .56 (N = 714) and .60 (N = 736) for the three successive quarters; and the correlations between total test scores and total performance scores for the same three terms were: .23, .18, and .34--considerably smaller than frequently reported relationships between test scores and ratings.

More Direct Empirical Validation

Studies focused upon attempts to more directly validate ratings subscales or items are less common. However, Morsh et. al. (1956) reported that "student ratings of instructors' subject matter knowledge was correlated significantly with instructors' proficiency test scores" (p. 88) although the correlation between rating and test score was only

0.19. McKeachie and Solomon (1958) and Sullivan and Skanes (1974) reported slight relationships between global ratings of instructor-course effectiveness and student entry into subsequent courses, but neither study included specific ratings items focused upon the student's intentions to take additional courses in the field or focused upon the teacher's elicitation of student interest in the field. Wood (Note 1) and Mihalo (1978) each compared student descriptions of the characteristics of 99 graduate teaching assistants with scores these teachers produced through the use of several well-known, standardized personality measures, and the resulting correlations were all insignificant or slight when signficant. Derry (1977) reported that the change of a course text caused a change in the text-related ratings item scores of subsequent classes without changing more global course or teacher ratings, but that the text ratings later dropped back to prechange levels even though the new text was still used. The original Frey (1973b) report in Science and his replications (Frey, 1976, 1978; Frey, et. al., 1975) also represent a relatively direct attempt to validate a specific factor of a student ratings scale, student self-reported accomplishment or learning gain, and although scores on this factor correlated with other indicators of student learning, Frey's (1978) more recent analysis of this and other factors presents the position that the "accomplishment" factor might better be interpreted as only one major component of the more global and inclusive dimension of "pedagogical skill".

Two questions included on many ratings scales ask students to report their expected course grade and their current grade point average. Dunnette (1952) found that reported grade point average correlated well (.94) with actual GPA for a group of college seniors; Richards and Lutz (1968) found high correlations (.84 and .87) between college-reported and student-reported GPA; and Walsh (1967) similarly found a strong relationship between student-reported expected grade and the grade actually received.

The use of lower-inference measures to concurrently validate the results obtained from the use of higher-inference rating items or subscales has been attempted through the use of observation scales originally developed for use with pre-college classes (Touq, Feldhusen, & Halstead, 1973), observation scales specifically developed for use with college classes (Love, Minnis, Sandoval, & Yonge, 1977; Love & Sandoval, 1978), or observer ratings devices designed to parallel a particular student ratings form (Centra, 1975), but the results have been mixed, even within each study. Many of the observation systems originally designed for use with pre-college instruction (see Simon & Boyer, 1970 for descriptions of many of the most commonly used procedures) frequently produce results characterized by such low generalizability across situations (Calkins, Borich, Pascone, Kugle, & Marston, 1977; Frick & Semmel, 1978; Shavelson & Dempsey-Atwood, 1976) that comparisons with ratings are difficult to interpret; Centra's (1975) use of colleagues as observers produced colleague ratings so high and, consequently so lacking in variability, that correlations with student ratings were similarly difficult to interpret; while the Love et. al. (1977) and the Love & Sandoval (1978) FLOAT observation system, although specifically intended for college use and producing some impressive subscale reliabilities, is still in its experimental/developmental phase. Even if

FLOAT or a similar, college-oriented, lower-inference, observational system is developed to a point where its use produces data as generalizable and/or as predictive of student learning (or other instructional outcomes) as the data now produced by the high-inference (but reliable) ratings scale, a lack of correlation between the two types of measures would justify the substitution of one measure for another only if one could be demonstrated to be more valid, less expensive and/or more useful in one or more functions. Even at the current level of development, FLOAT would appear to be capable of providing teachers with more specific feedback more potentially useful to a teacher that wished to change, and perhaps, improve classroom performance. An examination of the typical student ratings form will usually indicate that ease of scoring, ease of completion by untrained raters, and an associated interest in reducing the physical size (to permit optical scanning/processing) and/or length of rating forms have caused most developers to employ answers (cues) that are similar across questions (stems) and otherwise violate procedures designed to produce an independence of rater response across different items or subscales (see Sockloff, 1973, pp. 132-151 for specific criticisms of the construction of ratings forms).

FACTORS WHICH MAY INFLUENCE RATINGS: CONSTRUCT VALIDITY

Other Approaches to Validity

Much of the research focused upon the validity of student ratings has either: (1) attempted to determine if student ratings of teacher effectiveness can be substituted for student accomplishment as measured by course examinations; or (2) attempted to determine what course or situational characteristics--or teacher manipulations--are likely to cause ratings to move upward or downward. Since it is difficult, if not impossible, to compare student learning gains across different courses or different disciplines, it is probably impossible to directly evaluate teacher effectiveness, across courses, in terms of student learning. However, if student ratings can be demonstrated to parallel student learning gains within multi-section classes, such student ratings might validly be employed as a generalized substitute for student learning gains to compare different teachers against a common standard regardless of what courses are involved. Such a generalization of findings uncovered in large, multi-section, introductory courses to smaller, more advanced-level, college courses is unlikely ever to be tested directly in these smaller courses because most of these courses are only taught by one or two teachers at any one time. Consequently, it appears important to determine exactly which course or situational characteristics are likely to influence ratings in order to estimate the degree to which ratings do generalize across different courses and situations. If ratings are to be employed in such a normative function, it is also important to determine if teachers could "unfairly" manipulate particular aspects of their presentation or situation in order to gain a comparative advantage by increasing their ratings. A third focus of this type of validity study has involved the comparison of the results of student ratings with other criteria against which teachers are normally compared--usually against research productivity or against the judgments of peers or administrators. These comparisons are particularly indirect evidences of the validity of student ratings since the relationships between teaching and research are far from clear, and the validity of the judgments of persons who rarely observe the teaching of faculty is problematic at best.

Situational Factors that Influence Ratings

Conventional wisdom suggests that teachers who wish to increase their student ratings, without going to the effort required to improve their teaching and/or the achievement of their students, should: (1) promise and assign higher grades; (2) ask to be assigned to smaller, elective, higher level classes taken mostly by high GPA females who are majors in the field; (3) shift from the lecture to the discussion system or to the Keller-Plan; (4) carefully choose the items that are included on the ratings form; (5) tell the students that the ratings results will be used for administrative purposes, and that they must be signed by the students; (6) pass out the ratings form during the end-of-class party, immediately after the students have been lavishly praised for their hard work, dedication and extraordinary achievement; (7) remain in the classroom (sitting behind your pictures of spouse, children and dog) while students complete their forms; and (8) collect the forms yourself and return them to the department or college office--after having extracted those forms completed by those few students who have failed to recognize your abilities. If all else fails, complete the ratings forms yourself.

Each of the factors mentioned in this list of possible manipulations has been demonstrated to increase student ratings of teacher effectiveness, although there also are sets of data in which almost every one of these factors has been demonstrated to be unrelated to the student ratings of one or more groups of college faculty. Even when the preponderance of evidence supports a general and common relationship between a factor and ratings, it is usually impossible to determine from reported research, whether or not such a relationship is an evidence of a bias which partially invalidates a ratings score or an example of a logical relationship which provides evidence of the construct validity of student ratings. A series of thorough secondary analyses (Feldman, 1976a, 1976b, 1977, 1978, in press), the increasing use of multiple regression analyses, discriminant analyses, and canonical correlational analyses, and the accumulation of massive blocks of data derived from the use of the same instrument in scores of different settings (Cashin & Slawson, 1977a, 1977b; Centra & Creech, 1976) have been instrumental in the clarification of some of the relationships existing between ratings and other variables. However, the atheoretical nature of most of the reported research and the general inability of most experimenters to manipulate teacher characteristics, teacher assignment or the other major variables have produced a mass of data but few unambiguous answers.

Expected Grade

One of the most troublesome and most frequently investigated influences upon student ratings is the relationship that exists between the grade a student receives (or expects to receive) in a course and the rating (or "grade") given to the course or the teacher by the student.

Voeks and French (1960) asked, "are college students influenced by grades when they rate quality of teaching....do they assign high ratings to instructors who give high grades and lower ratings to instructors who give lower grades....if student ratings are little affected by grades, might they not be profitably used to evaluate teaching techniques" (p. 330)? After discovering that faculty members at the University of Washington "gave widely varying percentages of _A_'s and _B_'s and _D_'s and _E_'s" (p. 333-334), Voeks and French (1960) also concluded that "apparently, high ratings cannot be 'bought' by giving high grades, nor are they lost by giving low grades....When judging their instructor's over-all value as a teacher and when rating his skill in special respects...the students rarely, if ever, were influenced by the grades which they had received from that teacher" (p. 334).

According to Voeks and French (1960), their failure to discover a relationship between grades and ratings was evidence that "college students appear to have greater objectivity and less superficial value systems than we have realized" (p. 334).

The possibility that faculty could "buy" ratings by raising their grades, and that the "superficial value systems" of student raters would cause students to reflect their grades in their ratings, is a possibility that represents a continual worry for those interested in using student ratings as a comparative indicator of faculty effectiveness. An associated problem lies in the possibility that the institution of a ratings system or an increase in administrative pressure to improve ratings would accelerate a "grade inflation" already existant at many colleges.

Are ratings related to grades; does a ratings-grades relationship imply bias and invalidity; can faculty "buy" higher ratings by raising grades; and does increased use of student ratings lead to increased grade inflation? Feldman (1976a) reviewed over 45 studies of the ratings-grades relationship and Centra and Creech (1976) and Wood (Note 3) surveyed results derived from thousands of classes, and the answer to only the first of the four questions has emerged--although there appears to be considerable variation from teacher to teacher, class to class, and ratings item to ratings item, in general, students and classes that expect better grades also tend to produce higher ratings. Although there exist studies in which actual or expected grades are not positively correlated with student ratings (e.g., Blum, 1936; Heilman & Anmentrout, 1936; Overturf & Price, 1966; Sherman & Blackburn, 1975; and Voecks & French, 1960) the great majority of investigators have found correlations ranging from 0.10 to 0.44 between expected grade and student ratings of aspects of the course or instructor (see Feldman, 1976a, Tables I, II, & IV, pp. 71-80, 94-97 for a summary of 34 studies involving the pooling of students across classes and 11 studies using classes rather than students as the unit of analysis). Centra and Creech (1976) selected a random sample of approximately 15,000 students and 10,000 classes from their ratings data describing 300,000+ students, 16,000 classes and 100 post-secondary institutions. The correlations between expected grade and global ratings of teacher effectiveness and course value were .19 (N = 14,023 students) and .26 (N = 14,093 students) respectively when students were employed as a unit of analysis and .20 and .31 respectively for teacher and course when nearly 9,200 classes were employed as the unit of analysis (see Table 3, p. 25). The 4583 students expecting an "A" grade gave their teachers a global effectiveness rating of 3.95 (on a five point scale); those expecting a "B" grade (N = 6711) produced a rating of 3.74; those expecting a "C" (N = 2503) produced a 3.41 rating; while the 1.6% (N = 226) expecting a "D" produced a 3.02 rating that was 24% and nearly one standard deviation lower than that of their "A"-expecting classmates. All differences were significant beyond the .001 level, and the magnitude of this effect was considerably greater than those effects attributable to the students' GPA, class level or course electivity. Wood (Note 3) produced similar correlations between expected grade and global ratings when 35,000 students (r = .18 to .20) or 2,000 classes (r = .21 to .28) in a single university were employed as the units of analysis. However, different types of questions on Wood's scale produced somewhat different relationships with mean class expected grade: (1) items reflecting teacher scholarship, organization, clarity of presentation, effective use of class time and preparation for class were slightly related to expected grade (r = .06 to .13); (2) global ratings of the instructor (r = .20), the total course experience (r = .25) and student accomplishment (r = .28) fell into a middle range of relationship; and (3) items related to grading fairness and test quality (r = .35 to .36), items related to desire to take future courses in the field (.35) or from the instructor (.30), and items reflecting teacher-student (.34) or teacher-group interaction (.34), discussion quality (.31), or ability to individualize instruction (.36) were more strongly related to expected grade. One particular item, "Grade this course in comparison to your other courses this term ... A, B, C, D, F", produced a considerably higher relationship between its score and expected grade (r = .50 for 2089 classes and r = .39 for 35,257 individual students). A series of multiple regression analyses indicated that expected grade remained a better predictor of mean class scores on all types of ratings items even when such factors as teacher rank and experience, class size, course level and teacher attitude toward ratings and toward students' ability to rate were added into the prediction equations--although teacher experience indicators were nearly equal in the prediction of scores from the scholarship, lecture clarity and organization items. The calculation of omega squared and eta squared statistics indicated that while expected grade accounted for less than 2% of the variance of scores from questions relating to lecture ability and from 4 to 5% of the variance of scores from global questions, expected grade accounted for from 6 to 10% of the score variance of questions more directly related to student attitude toward the course or teacher, toward the teacher's ability to develop rapport, promote discussion or grade with fairness, and for over 15% of the score variance of a question asking students to grade the course. An oddity of the data that was also mentioned in Feldman's (1976a) review was that students expecting to fail did not produce lower ratings than students expecting a "D" grade or even a "C" grade.

Velicer (1977) has also concluded that expected grade is less likely to be related to items dealing with how material is communicated and more likely to be related to items involving information about the students' appreciation of course products, student appreciation of the course, or student opinions concerning evaluation procedures. Elmore and Pohlmann (1978) employed zero-order correlational analyses and a canonical correlational analyses to ratings and to the student/teacher/class characteristics of 174 classes that used the same 20-item ratings scale. As was true in the Centra and Creech (1976) and Wood (Note 3) results, expected grade was the most potent correlate of ratings. As was true for the Wood (Note 3) results, expected grade correlated more with scores from items reflecting rapport and appreciation (showed an interest in students, .49; increased your appreciation for the subject .40) or openness to discussion and student participation (encouraged student participation, .49; knew if students understood him or her, .45; accepted criticism and suggestions, .39), than with scores from items reflecting the teacher's preparation for class (r = .13), clarity of assignments (.12), or apparent knowledge of the subject matter (.11). The most global question, "In general, taught the class effectively" (see Elmore & Pohlmann, 1978, Table 1, p. 188 for a listing of scale items and zero-order correlations) correlated moderately (r = .31) with mean class expected grade. Only the set of canonical functions representing the largest canonical correlation (.83) was reported by Elmore and Pohlmann (1978, Table 2, p. 190), but it was dominated on one side by mean expected grade (.81 loading) and on the ratings side by items such as "encouraged student participation" (.76 loading), "showed an interest in students" (.75 loading) and "knew if students understood..." (.65 loading). Potential biasing characteristics which also loaded on the non-ratings side of the largest canonical relationship were: class size (-.56 loading); the instructor's self-ratings of personal warmth (.55); mean student year in college (.46); mean class GPA (.43); and the instructor's self-rating (.43) of primary teaching focus as the course content (scored as "0") or the students (scored as 1).

Frey (1978) factor analysed data from the use of his seven-item ratings scale with 1298 undergraduate classes and 26,787 students at

Northwestern University. Two global factors accounted for over 75% of the variance, and Frey weighted different combinations of his 7 ratings items to produce ratings scores that reflected these two factors. Factor 1 was composed mostly of the scores from three items relating to the carefulness of class planning, presentation clarity and student gains in knowledge and competence, and was labeled "pedagogical skill". The second factor reflected the items relating to welcomeness of class discussion and availability of personal help, and to a lesser degree, the accuracy of the grading procedure, and was labeled "rapport". The rapport factor correlated well with mean class grade ($\underline{r}$ = .46) and with class size ($\underline{r}$ = -.40) while the skill factor was poorly correlated with mean class grades ($\underline{r}$ = .02) and with class size ($\underline{r}$ = .07). Analyses of variance indicated that class size and mean class grade had effects of similar magnitude upon rapport ratings but that the two class characteristics did not interact. A reexamination of previously reported relationships between student ratings and the achievement test scores of multiply-taught, multisection courses (Frey, 1973, 1976; Frey, Leonard & Beatty, 1975) permitted Frey to calculate the relationships between the rapport and skill ratings factors and the mean class test scores that reflected course achievement. The skill vs. test correlation coefficients for the eight multisection courses ranged from .99 to .26 with a median correlation of .81 and an unweighted mean correlation of .77--while the corresponding rapport vs. test correlations ranged from .05 to .82 with a median correlation of .29 and an unweighted mean correlation of .36 (see Frey, 1978, Table 3, p. 86).

The frequently-reported relationships existing between scores from one or more types of ratings items and the grades students receive/or expect to receive) present administrative and theoretical problems and questions. The belief that increasing students' grades will increase global ratings (or ratings of rapport) could possibly lead to a grade inflation for those faculty most pressed to increase their ratings or most wishing to increase the apparent satisfaction of their students--even if such a belief were not valid. Although Centra and Creech (1976, p. 27) stated that "there is no evidence, however, that ratings programs have indeed been a major contributor to grade inflation", they report no study proving the contrary or even any study in which the question is examined. Another administrative question involves the suggestion made by Frey (1978, p. 88) that "rapport and skill should be measured separately and that evaluation decisions should be based primarily on the skill factor" in order to protect faculty teaching large classes or grading conservatively from being placed at a "competitive disadvantage with their academic colleagues". Should reflectors of student rapport and positive student attitudes and those teacher or class characteristics more associated with discussion than with lecture be excluded from ratings instruments used to gather data for normative purposes? Should the rapport, attitude or discussion-related questions be included on ratings questionnaires while the results are adjusted for mean class grade or class size or class style?

The more theoretical questions are more focused upon the meanings underlying the apparent relationships between ratings and grades. Why are strong relationships found for some instructors but not for others (see Feldman, 1976, Table III, p. 81 for an example) who teach the same courses and use the same ratings questionnaires? Although the instability attributable to small sample sizes presents obvious problems, there has been little or no investigation of the factors that may produce ratings-grades relationships even though this type of nearly atheoretical, empirical investigation is the type of ratings research most frequently reported. A more difficult question asks if the reports of the ratings-grade relationships are evidences of ratings bias or evidences supporting the validity of ratings. If at least some ratings factors are more related to affective outcomes than to student gain of those lower-level cognitive skills most represented on course examinations--a relationship not unequivocally demonstrated, and if course grading policy is also related to affective outcomes, is a relationship between those rapport-related ratings items and mean class grades more an evidence of the bias of the scores produced by these ratings items or the validity of those scores. Frey (1978) incorporated student assessment of their own achievement into his skill factor and implied that the lack of grade-"bias" found for the skill factor also applied to the scores on the self-assessment of achievement item. However, Wood (Note 3) found a significant ($\underline{r}$ = .28) and noticeable correlationship between mean course grade and a similar accomplishment item, and found the accomplishment item to be very similar in distribution and relationships to the global items of teacher and course effectiveness. The relationships between class grades and scores on items reflecting effective class discussion, faculty-student interaction and faculty self-reported attitude toward students (Elmore & Pohlmann, 1978) suggest that effective promotion of discussion or other student-centered types of instruction (styles believed by McKeachie, 1967 to be effective in promoting higher level cognitive objectives) may possibly produce higher grades as faculty identify these hard-to-measure gains in their students or as the more student-focused faculty find it difficult to potentially "penalize" their students with lower grades. These possible explanations for the grade-ratings relationship imply that higher grades do not produce higher ratings, but that factors related to both (increased higher-level learning and/or faculty orientation and/or teaching style) cause the grades-rating relationship.

The generally atheoretical nature of most ratings research and the consequent general lack of data reflecting independent assessments of student-teacher-class characteristics and their interactions presently preclude any determination of whether or not ratings are biased by expected or actual grade distributions--or whether or not an instructor will improve ratings by assigning higher grades.

Grade Point Average (GPA) and Grade Discrepancy

None of the data reported by Bausell and Magoon (1972), Centra and Creech (1976), Elmore and Pohlmann (1978), Feldman (1976a) or Wood (Note 3) suggest that students with higher cumulative grade point averages rate their teachers or courses much differently than do those students with lower self-reported GPA. Centra and Creech (1976) found no significant difference between ratings by students of different grade points but did find (as did Bausell & Magoon, 1972) that students with "A" or "B" grades who expected a "C" or a "D" in the class did produce slightly, but significantly, lower ratings. Wood (Note 3) similarly found that, when nearly 30,000 students were the unit of analysis, GPA tended to mirror the relationship between expected grade and a ratings item, but produced a correlation coefficient only a fifth to a tenth as large while the grade discrepancy indicator ((GPA + 5) - Expected Grade) also mirrored the expected grade but at about half its size. When whole classes were employed as the unit of analysis, correlations between GPA and an item score tended to be about half the size of correlations between expected grade and the item score while the correlations between discrepancy score and ratings item scores were only about a fourth the size of those between expected grade and ratings scores (and half of those between GPA and ratings scores). In the Wood data (Note 3), at least part of the relationship between ratings and GPA, limited as it was, can be attributed to the fact that grade point average is the best single predictor of expected grade ($\underline{r} = .4$).

Holmes (1972) created one of the few experimental investigations in the ratings area when he randomly selected 48 of his 97 A or B students (all of whom were expecting the grade they eventually earned) and informed them that, due to a poor performance on the last of four course examinations, they would actually receive a grade one level lower than expected. Holmes (1972) reported that there was no difference in ratings attributable to expected grade but that the ratings scores of typical lecture or skill items (instructor was well prepared for class; material was presented coherently; examples clarified material) were depressed for the group experimentally assigned to the lower-grade group. Scores on items reflecting student stimulation were also depressed for the experimental group, but scores on the typical discussion-rapport items (respect for students; freedom to ask questions) were not depressed. Holmes concluded that disconfirmed grade expectancies did bias ratings of teaching performance but that such disconfirmation of expectancies should not be typical of situations in which students are kept aware of their progress and in which grades accurately reflect proficiency. Whether Holmes' experimental manipulation really influenced skill-lecture ratings and failed to influence rapport-discussion ratings cannot be proven from the data reported by Holmes, but if true, why weren't the rapport-discussion ratings influenced? Could the large size of the class have made only the ratings of lecture skills relevant and hence manipulatable? Would the results found in such an experiment generalize to smaller classes or to classes in which students were less able to predict their final grade or to classes without experimental manipulation of grades?

In conclusion, although grade point average or grade discrepancies have been linked with ratings results, the direct relationships tend to be so slight that they are most likely to reach significance only with large samples or in unusual situations. For example, the global ratings of students expecting a course grade of "D" were 24% lower than those expecting an "A" from the Centra and Creech (1976) data and 24½% lower in the Wood (Note 3) data, but corresponding reductions in global ratings between students having an "A" level GPA and students having a "D" level GPA were 2% in the Centra and Creech (1976) data and 2.8% in the Wood (Note 3) data. Centra and Creech (1976, Table 5, p. 30) also provided comparable data for estimating discrepancy effects: "A" students who expected "C" or "D" grades produced ratings from 5 to 8 percent lower than "C" or "D" students expecting C or D grades; "A" students expecting "B" or "C" course grades similarly produced global ratings from 5% to 7% lower than "B" students expecting "B" or "C" grades; but all students expecting an "A" grade, regardless of their

GPA, produced similar ratings. Wood (Note 3) similarly reported that grade discrepancies accounted for only 2% of the variance of global ratings (eta squared = .018) as compared with the 4 to 5% associated with expected grade and the 0.08% associated with student GPA.

Class Size

Centra and Creech (1976) and the researchers with access to the other massive, multi-university set of ratings data (Cashin & Slawson, 1977a, 1977b; Hoyt & Cashin, 1977) have each identified class size as one of a few variables most associated with student ratings, although Wood (Note 3) reported that class size was only slightly related to the rating of nearly 40,000 students at one college. Feldman's (1978) review of course characteristics and student ratings caused him to conclude that two-thirds of the nearly thirty studies producing product-moment correlations found "indications of a negative relationship--the smaller the size of the class, the higher the ratings...thus the variable of size in these studies usually explains somewhere between 1% and 7% or 8% of the variance of class ratings, depending on the particular rating item and the population under study" (p. 206). The general pattern of results produced by the various class-size studies suggest that at least part of the relationships between class-size and ratings are likely to depend upon the teacher-assignment and student assignment policies of the various departments and colleges. If superior lecturers are automatically assigned to teach the large introductory courses or if popular teachers are permitted or encouraged to draw as many students as possible into their classes, several, or even all, of the larger classes should elicit high ratings. In situations where class size is controlled and overflow enrollments are automatically assigned to other sections taught by other teachers, or where the more effective lecturers are not assigned to the larger classes, the true effects of class size, independent of teacher effects, are more likely to be exhibited. Only one set of studies (Bausell, no date; Bausell & Vinograd, 1977; Holland, 1954; Remmers, Hadley & Long, 1932) controlled for the teacher effect by comparing ratings across class sizes within the set of classes taught by each teacher. Bausell (no date) identified 130 pairs of courses that were taught by the same instructor but that differed in enrollment (one class was at least 1.5 times the size of the other). An analysis of variance indicated that the class size effect was significant ($\underline{p} < .001$, $\underline{f} = 28.2$, $\underline{df} = 1, 127$) even when course level had been partialled out. A post-hoc analysis of 16 pairs of different sized sections of the same course taught by the same teacher "revealed that in 13 of the 16 cases the smaller course received a higher rating than the larger, there were two ties, and in only one case did the larger receive the higher rating" (p. 2). The Bausell and Vinograd (1977) study was a larger version of the Bausell (no date) study and involved 254 pairs of courses, 26 pairings of similar sections, and produced very similar findings. This larger study indicated that instructors accounted for 31% of the ratings variance but that class size appeared to account for an additional 10% of the ratings variance.

The relationships existing among different ratings items, different class characteristics and class size were not explored by Bausell (no date) or Bausell and Vinograd (1977), and the Holland (1954) and Remmers, Hadley & Long (1932) studies include too few cases to permit such analyses. Although Frey (1978) examined the effects of class size across, rather than within instructors, his analysis of 1298 classes at Northwestern University caused him to conclude that "there was a strong, negative relationship between class size and ratings of rapport....the product-moment correlation was -.40....the largest drop in ratings on this factor seems to occur when the enrollment increases from 30 to 50 students....with the students perceiving dramatically less classroom interaction when the enrollment exceeds this value" (p. 78). Frey also reported that his "pedagogical skill" factor was weakly but positively related to class size (r = .07) with the larger classes producing slightly high ratings on this factor. Frey also reported that there were no interactions between class size, mean class expected grade, and instructor experience on ratings of rapport, but that the larger classes (41+) were more frequently taught by tenured instructors who received higher ratings on pedagogical skill than their untenured colleagues who less frequently taught larger classes. Cashin and Slawson (1977b) reported a set of 67 tables representing scores on all items of the IDEA ratings system broken into five levels of student motivation and four levels of class size (1-14, 15-34, 35-99, 100+) and representing 17,234 classes from over 100 institutions (see pp. 46-65). According to Cashin and Slawson (1977b) "as a general rule...the smaller the class, the higher their ratings....the one instance where there may truely be an exception to the 'general rule' about size is with the Very Large classes....in many instances, Very Large classes receive higher ratings than Large classes" (p. 47). While the "promoted discussion" item score dropped from 4.344 to 4.233 to 4.0911 to 3.707 across the four class sizes for a 14.6% loss from the smallest to the largest class size, a typical lecture or skill item "spoke with expressiveness and variety" produced a similar total drop across size (4.359 to 4.342 to 4.303 to 3.807) for a 12.7% maximum drop from the smallest to the largest class size--but the drops across the three smaller class sizes were only from .3% to 1%. A similar small (less than 1%) drop across the first three size categories was reported for other lecture-related items: "demonstrated the significance of the subject"; "seemed enthusiastic about the subject matter"; "made dry and dull presentations"; "summarized in ways which aided retention"; "explained course material clearly"; and "related material to real life situations".

Centra and Creech (1976) presented data concerning 4760 classes that were described by their instructor as lecture-discussion classes, 876 classes described as lecture courses and 698 classes described as discussion classes. All were randomly drawn from a population of 16,000 classes (and 300,000+ students at 100+ institutions) using the SIR rating form between September 1973 and December 1974. Only the scores from the global, overall teacher effectiveness item and the "value of the course" item were examined, but Centra and Creech (1976) concluded that: (1) smaller lecture-discussion classes "were clearly rated higher than larger classes...in particular, classes with fewer than 10 or 15 students received significantly better ratings" (p. 51); (2) rating and class size appear to be related in a curvilinear fashion with classes of over 100 students increasing in rating and being rated similar to those in the 16-35 student range" (p. 51); (3) classes in the 35 to 100 student range received the lowest ratings" (p. 51); (4) the sharp drop-off in ratings occured around the 35+ class-size for lecture classes but there was a slight drop for teacher effectiveness ratings and a sharper drop for "value of the course" beyond a size of 10 for discussion classes and a greater drop (7%) beyond the 15+ size for discussion classes. The highest ratings for course value were achieved by discussion classes with 10 or fewer students. The authors concluded that if ratings were used as a criterion of optimal class size, "discussion courses with up to 15 or so students seem appropriate, while lecture courses would have up to 35 or so students" (Centra & Creech, 1976, p. 53). Several other researchers (Delaney & Kojaku, 1977; Gage, 1961; Haslett, 1976; Marsh, 1976; Pohlmann, 1975; Wood, Linsky & Straus, 1974) also found a similar curvilinear relationship between ratings and class-size.

Wood (Note 3) found significant but non-linear and quite small differences across classes of different size on ratings items related to lecture skills and presentation clarity; linear and small to moderate (eta square = .02 to .03) relationships between class size and ratings of self-reported learning; and linear and slightly larger (eta square = .05, _r_ = -.21) associations between class-size and ratings of attitudes toward the class and teacher (would like to take again).

Cashin and Slawson (1977a, 1977b) and Hoyt and Cashin (1977) consider class size a sufficiently important consideration to justify comparing teachers only to other teachers of similar size classes, and Frey (1978) has suggested that only pedagogical skill items be included in ratings used for normative purposes. Whether a relationship between ratings and class size represents bias is still a moot point and will remain so until the major factors, the teacher, the teacher-course interaction, the teacher preference for particular types of instructional modes, the course, course level and expected grade are better controlled or explained--and until particular and independently measured course inputs and outcomes are employed according to some theory or theories of instruction.

Recent reanalysis of some 300 reports dealing with the effects of class size on the achievement of elementary and secondary school students (Cahen & Filby, 1979; Glass & Smith, 1978) have presented the position that when class size is reduced to below 15 students, achievement test scores rise dramaticly. Students exhibiting average (50th percentile) levels of achievement in classes including 30 to 40 students might be expected to achieve at the 55th percentile in 20 student classes, at the 65th percentile in 10 student classes and at or above the 75th percentile with student-teacher ratios of 5 or less. Although there are small numbers of one to 15 student classes at most educational levels, the apparently parallel nature of the Glass-Smith Curve of Achievement and Class Size (Cahen & Filby, 1979, p. 493) and Centra & Creech's (1976) plots of class size vs. student ratings of teacher effectiveness (pp. 50, 52) provide an indirect support for the complementarity of student ratings of teacher effectiveness and student achievement. Cahen and Filby (1979) are presently attempting to experimentally manipulate class size (in selected elementary schools) in order to examine many of the resultant shifts in teaching procedures and in various student outcomes. Similar experiments would be even more easily organized at the college level in multi-section classes.

Student Interest

Student motivation to take the course or interact with the instructor is the variable other than class size that Hoyt and Cashin (1977) felt was necessary to control when their IDEA system was constructed, and the tables presented by Cashin and Slawson (1977b, pp. 46-65) indicate that student response to the question, "I had a strong desire to take this course" indicate that responses to this question are positively related to the responses to other rating items. Hoyt and Cashin (1977, Table 10, pp. 23-25) also presented evidence that responses to this item correlated positively with student ratings of progress for 1439 trial classes. Typical correlations between this interest score and progress ratings were: .43, .56, .62 and .66 for correlations between interest and progress in gaining factual knowledge in small, medium, large and very large classes respectively; .50, .53, .59, .59 relating to progress in learning principles and theories in classes of the same four size categories; and .57, .58, .69, .71 for progress in "discovering the implications of the course materials for understanding myself". The mean (unweighted) correlation coefficient between the ten progress scores and the interest score, for very large classes was .57, and for small, medium and large classes, was, respectively, .40, .50, and .59. Apparently, post-class, self-reported interest in taking the course could account for a proportion of ratings variance ranging from 16% to nearly 50% depending upon the ratings item and the class size. Whether or not this relationship between student interest and ratings of progress is independent from the effects produced by different teachers has not been determined although the use of student interest as one of two control variables in the IDEA rating system (Hoyt & Cashin, 1977) apparently was based on the supposition that no variance attributable to student interest was shared with that attributable to teacher effectiveness. Another question involves the use of a question embedded in a ratings scale and sharing a set of common cues with adjoining ratings items as a measure of an interest that existed before the class began and several months before the question was asked. Gillmore (1975), Gillmore and Naccarato (1975), Marsh (1978), and Marsh, Overall and Thomas (1976), produced similar but considerably weaker associations ($\underline{r}$ = .10 to .24) between ratings and student's self-reported interest in the subject matter before enrolling in the course, leading to the possible interpretation that some of the larger correlations found in the Idea results may be attributable to a "halo" effect or to a misreading of the question as asking for current attitudes.

Several researchers have attempted to obtain a somewhat similar measure of interest in a course by asking students to categorize the course as either an elective or a required course. In some cases students who identify the course as an elective rate it higher (Downie, 1952; Evans, 1969; Gage, 1961; Lovell & Haner, 1955), but in other cases there is little or no difference (Heilman & Anmentrout, 1936; Remmers, 1929). Similar studies have used the class as the unit of analysis and reported positive correlations between ratings and percentages of students electing the course (Aleamoni & Thomas, 1977; Landis & Pirro, 1977; Marsh, et. al., 1976; Pohlmann, 1972, 1975) or no particular pattern of positive correlations between course electiveness and ratings (Elmore & Pohlmann, 1978). Centra and Creech (1976) expanded the elective-required categories to better reflect the reason the student enrolled in the course by employing two separate questions. The first question asked if the course were: (1) a major requirement (or elective in a major field); (2) a minor requirement (or required elective outside the major field); (3) a college requirement (not part of a major or minor); (4) an elective not required in any way; or (5) other. The students responding with "elective-not required in any way" and "major requirement or elective in major field" produced the highest ratings of overall teacher effectiveness (3.77), with the "college requirement but not part of major or minor" students producing a significantly (<.001) lower global rating of the teacher (3.63 or only 3.7% lower than 3.77). Centra and Creech obtained a more noticeable difference when they asked students why they took the course. The 50% who indicated that the course was required produced a global teacher rating (of 3.59) which was similar to the (3.63) rating of those labeling their course as a college requirement. Those (30%) who chose the course because "the subject was of interest" produced a global teacher rating that was 7% higher (3.87) but similar to those choosing the course because of the recommendations of friends (3.83) or of advisors (3.80). However, those choosing the course because of the teacher's excellent reputation (only 5.3% of the students) produced a very high (4.64) global evaluation of the teacher's effectiveness that was 23% and approximately one standard deviation above that produced by those who saw themselves as taking a required course. For unknown reasons, those students (fewer than 2%) who reported that they took the course to "make a good grade" produced the lowest teacher ratings of all (3.15) --ratings significantly and noticeably lower than any other group of students.

As with research dealing with other possible biasing factors, unless the researcher is able to control or explain the major sources of ratings variance--most notably those associated with the instructor, the course and the instructor-course interaction, or unless the investigator is able to manipulate a major variable within sets of instructor-course combinations, the relationships between ratings and potential biasing factors remains unclear. Marsh (1978) reported that student-reported reason for taking the course was related to ratings even after expected grade and other factors were statistically partialled out, and Centra and Creech (1976, Table A-2, p. 70) also provided evidence that course electivity and/or association with major was still a significant influence upon ratings ($\underline{F} = 27$; $\underline{p} < .001$; $\underline{df} = 1, 11{,}779$; but omega squared $< .003$ and eta squared $< .003$) although a small one, and that there was an interaction between electivity/majorness and course level. On the other hand, Mirus (1973) and Shingles (1977) produced insignificant or mixed relationships between ratings and electivity when other factors were controlled, and Bausell and Vinograd (1977), the only researchers to control for teacher effects, while finding additional effects attributable to class size, found no additional ratings variance attributable to electivity, course level or instructional mode.

In conclusion, although student motivation to enroll in a course or for an instructor appears to predict or correlate with student ratings, course electivity or position of the course within or outside of the major sequence appears to be less strongly and less obviously related to student interest or to student ratings.

Subject Matter

Feldman (1978, Table 2, pp. 219-224) reviewed 24 studies which produced information concerning the relationship between academic area and ratings and concluded that: (1) none of the studies provided adequate and simultaneous control for factors which might vary across academic areas and which might provide alternate explanations for any area-ratings relationships; (2) differences in ratings of teachers in different academic areas tended to appear in those studies which did employ some controls for other factors; and (3) the eleven studies reporting ratings across three or more academic departments were neither completely comparable nor consistent in their results, but there was a tendency for teachers in the humanities (English, arts, humanities, languages) to be rated well, and for those in non-biological natural sciences, mathematics and engineering to be rated lower--with the social science teachers (political science, sociology, psychology, economics) to be in the middle. Feldman suggested that differences in mean class expected grade and sex distribution within a class are logical factors to be controlled in future studies as these are factors that may differ across academic areas. Batchelder and Keane (1977) reported that at their college, instructor lecture was the dominant style of classroom interaction in the physical sciences, somewhat less common in the social sciences, and least used (although still the most typical mode) in the humanities. Since Centra and Creech (1976, Table 13, p. 47) reported that the lecture with "little or no discussion" was associated with lower ratings, it is possible that some of the possible ratings differences across academic areas may be attributable to the instructional methods differentially characteristic of different academic areas. On the other hand, the Batchelder and Keane findings may not generalize across institutions, and even if they did, association of lecture with the lower-rated physical sciences could as easily mean that this association produced lower ratings for the lecture method instead of the use of the lecture method leading to lower ratings.

Method of Instruction

Durbin and Taveggia (1968) reviewed more than 350 studies of college teaching and concluded that there were no differences in student learning attributable to the use of any of the various face-to-face instructional systems (lecture, discussion, laboratory, etc.) or the use of televised instruction or independent study as a substitute for or complement to the more commonly used face-to-face instructional methods. This survey reported results similar to those reported in previous surveys (McKeachie, 1963; Nachman & Opochinsky, 1958; Wallen & Travers, 1963). There has been no comparable body of literature focused upon the relationship between teaching method and student ratings or teaching method and affective outcomes although Centra and Creech (1976) have reported that courses described by their instructors as "lecture, little or no discussion" produced somewhat lower (5%) global teacher ratings (3.64) than those produced by "discussion mainly" (3.85) or "lecture and laboratory" (3.81) or "lecture and discussion combined" (3.76). All differences were statistically significant but slight. Since Centra and Creech (1976) deleted instructional method from their subsequent multifactor analysis of variance, they provided no evidence of this relationship controlled for other factors. Although the IDEA system asks instructors to provide similar descriptions of their instructional method, the first three IDEA technical reports (Cashin & Slawson, 1977a, 1977b; Hoyt & Cashin, 1977) do not provide data concerning relationships between methods and ratings.

Since 1968, various versions of the "Keller System" (Keller, 1968) and other teaching methods related to "mastery learning" (Block & Anderson, 1975; Bloom, 1974, 1976) or "behavioral instruction" (see Robin, 1976 for a review of the literature relating to college teaching) have produced evidence that the application of various versions of these product-oriented or examination/content-oriented instructional systems can produce higher levels of student learning as well as higher student ratings than produced by the more typical lecture or lecture-discussion approach. Although Doyle (1977) has likened these systems of instruction to the tutorial approach because of the common use of numbers of course aides or proctors, there is little evidence that proctors actually function in instructional rather than only evaluative capacities. Badia (Note 4) has compared the effect upon both achievement and ratings produced by five sections of a psychology class taught by the Keller method (weekly mastery tests with opportunities to retake alternate versions in order to achieve a preset level of success, proctors, home study replacing lectures) and thirteen lecture sections that differed from one another only in terms of test frequency (3 tests, 5 tests or 9 tests during the term). Since both ratings and student achievement (on the same 165 achievement test items) increased with the increase in test frequency, an obvious possibility is that, up to some optimal frequency, both ratings and achievement are dependent upon a factor that can transcend method--test frequency. Both Badia (Note 4) and Wood and Bergman (Note 5) reported that college students appear to prefer a testing schedule that involves biweekly testing, but the Badia (Note 4) results may be interpreted to mean that a weekly or slightly more frequent testing program may produce higher achievement as well as higher ratings. Although Badia's grading system for his Keller classes (and also for his lecture classes) was entirely dependent upon student score on a common set of 165 typically normative achievement test items, it is not uncommon for other Keller-plan or mastery classes to base the student grade upon eventual completion of a series of short tests which may be retaken (in alternative versions) more than once. Such a procedure usually increases both the mean student grade and the amount of student effort--both potential correlates of student ratings, without necessarily producing an equal increase in student achievement. Consequently, increased expected grade and increased student effort may provide additional causes for the increased ratings often associated with the Keller and Mastery methods. Some of the research related to mastery or Keller systems has indicated that procedures which increase student achievement (e.g. required remediation, Kulik, Kulik & Hertzler, 1977; obligatory contracting for grades, Burton, Pathak, & Burton, 1977) may not increase student ratings while procedures which increase student ratings (e.g. the use of student proctors, Arp & Semb, 1977) may not increase student achievement. One possible explanation for the Burton, Pathak and Burton (1977) reduction in ratings may have been caused by their possibly punitive grading system which provided for strong penalties for students unable to meet the obligations associated with their chosen grade level. For example, students unable to meet the requirements of their "C" contract but able to meet the requirements of a "D" contract would not receive a "D" grade but would have this lapse "result in automatic failure" (p. 130). Burton, et. al. (1977) state that "an element of uncertainty, intended to further motivate performance, was introduced as follows: (a) no contract level could be lowered once the commitment was made; (b) failure to complete the assigned quantity of work would result in automatic failure...(p. 130), but they also attributed their lowered ratings of "every facet of the course" (p. 132) to the experimental and innovative nature of their contracting technique rather than to their anxiety-producing and probably punitive grading system.

Although much of the productive, process-product research may be occurring within the context of the Keller or mastery-learning paradigms, an empirical study, by McMillan (1977), was outside this context and produced data indicating that a positive attitude toward a segment of a course were generated by teacher praise (vs. no feedback) and by a high effort assignment (writing a three to four page answer to a question related to the reading assignment vs. writing a one page answer). The author also reported an interaction between praise and effort. Apparently a successful identification of the method-ratings-achievement relationships are less like to be discovered by examining the effects of major categories of teaching methods (lecture vs. discussion vs. TV) and more likely to be found in a more microscopic examination of class procedures—most probably those relating to testing, grading, monitoring and feedback processes.

Course Level

Some of the research related to course level has already been reported in the section dealing with the relationships between student interest and ratings. Most notably, Bausell and Vinograd (1977) found that there was no relationship between course level and ratings when class size and the teacher factor were controlled, and Centra and Creech (1976) found a small interaction between elicitivity and course level. Since student level and/or course level are very likely to covary with expected grade, class size, the electivity of the class, the likelihood that the course is in the major, and familiarity of the student with the reputations of teachers and courses, a simultaneous inspection of these covariables would be necessary for the determination of the effect attributable to course level itself. Although student level (freshman, sophomore, *etc*.) or course level (introductory, upper level undergraduate,

graduate, etc.) are descriptors that are easily obtained, it is not clear as to what information is likely to be conveyed by this factor that is independent of possibly related factors such as expected grade, student familiarity with reputations, etc. Centra and Creech (1976) found no relationship between the level of the course and global teacher ratings when the courses were classified by 8637 instructors (see Table 7, p. 33), and very little difference in the ratings made by 14,236 students who were categorized as to their class level (although the juniors produced ratings approximately 2% lower than those of graduate students or other undergraduates). Feldman (1978) reported finding 21 studies that found small but positive zero-order correlations between course or student level and ratings, seven studies finding no zero-order relationship, five finding a mixed pattern of relationships depending upon the ratings item, two multivariate studies finding a positive relationship after other factors were controlled, seven finding no such relationship, and one multivariate study producing mixed relationships. None of the positive relationships were very strong.

Student Gender

Centra and Creech (1976) found no differences between the ratings produced by male and female students, and this finding replicates results of the smaller studies by Delaney, 1976; Granzin and Painter, 1973; Remmers and Elliot, 1948; and Spencer, 1965. However, Aleamoni and Thomas, 1977; and Elmore and Pohlmann, 1978, have reported small differences indicating that women may produce slightly higher ratings. Several studies (Elmore & LaPoint, 1975; Feild, Holley & Armenakis, 1976;

McKeachie & Lin, 1971; and Walker, 1969) have found that female students may rate female teachers higher than male teachers or that male students may value structure more than female students (who may value teacher warmth more).

The Ratings Forms

Can teachers obtain higher or lower ratings by including particular items on their ratings forms or by excluding other items? Ranney (1977) employed the Purdue CAFETERIA instructional evaluation system to create a small experiment to test this point. Seven teachers who each taught two similar sections of a course were each asked to construct two, five-item rating forms by selecting five items (of the 200 available in the Cafeteria system) that each individual teacher believed would elicit high ratings and five other items that were believed to elicit less favorable ratings. Each of these two scales for each teacher was expanded through the addition of the six items used by the system to produce global, normative ratings of the teacher/course. One such scale was given to each class in each pair of classes. Ranney's analysis of the results led him (her?) to conclude that: (1) the instructor-chosen sets of five "favorable" items produced ratings from each of its classes that were higher than those produced by the paired classes on the "unfavorable items"; but (2) the responses to the six global items were similar across paired classes, uninfluenced by the five teacher-chosen items, and different for the different teachers. If Ranney's small study is generalizeable to other classes and teachers, it provides evidence that teachers can influence their ratings through careful selection of items; that teachers tend to be able to identify items that reflect their strengths and weaknesses; and that student responses to global items may be relatively free from being influenced by more specific items that precede them on a rating form. Each of these assumptions are necessary prerequisites to the fair use of a Cafeteria system which permits faculty to select the more specific, Level I items from an extended menu, but automatically follows these items with a common core of global, Level II items that may be employed for normative/comparative/personnel purposes.

A series of studies by Bendig (1952), Follman, Lucoff, Small and Power (1974), Leftwich and Remmers (1962) and Sharon and Bartlett (1969) have demonstrated that changing the response format of ratings items is likely to modify both the mean and skewness of the distribution of responses. However, a similar series of studies by Follman, Lavely Silverman, and Merica (1974), Follman, Merica, and Silverman (1974a, 1974b), Grasha (1975), Jaeger and Freijo (1974), and Remmers and Brandenburg (1927) have produced evidence indicating that the item stems can be amplified by more explicitly defining traits or comparison groups and that the ordering of the cues from positive to negative can be reversed without greatly effecting student responses.

A body of experimentation has focused upon the written directions provided to the student at the top of the ratings form--most notably directions indicating if the ratings results may be used for personnel purposes or for instructor improvement or research purposes. Feldman (in press, pp. 5-8) has reviewed 9 such studies and has concluded that: (1) there is a general tendency to find that students produce somewhat higher ratings when they are informed that the results may be used for personnel or other normative or administrative purposes; (2) this effect may vary from teacher to teacher; and (3) the relationships are not very clear. Pasen, Frey, Menges, and Rath (1978) increased the power of this effect by having the department chairman read the administrative version of the directions to a segment of 2 classes while the instructors read the "for my own use" directions to other students. While "pedogagical skill" ratings were only slightly increased by the administrative direction, the rapport ratings were increased by nearly one standard deviation.

In summation, although major modifications of ratings item stems may produce little difference in scores, the use of different items, different indications of the purpose of the ratings, and the use of different response keys may all influence the ratings scores in the favor of those teachers who select their own items, claim that the ratings results will be employed for administrative purposes, and employ graphic response formats, the strongly agree, agree, etc. response format or unbalanced sets of response cues. Some evidence exists that indicates that common global items may not be influenced by the more specific questions that precede them in a ratings form.

Administration of the Ratings Form

Factors related to form construction and form administration are more easily manipulated than factors such as expected grade or teacher distribution across different types of classes or students. Consequently, experiments tend to be more common in this area of research than in most other areas of ratings research. One such experiment was conducted by Fentress and Swanson (1975) as part of a master's degree project, and involved twelve instructors (all graduate teaching fellows), and their twelve classes (204 students, 85% male). During the first academic quarter of the experiment, intact classes were randomly divided into two similar groups, 12 experimental half-classes and 12 control half-classes, and each half-class met separately to take the final examination and to complete the Illinois CEQ rating form (Spencer & Aleamoni, 1968). The 12 teachers were provided with a list of eight statements (e.g. This has been one of my most productive classes; This is one of the best classes I have ever had; Your work has shown steady improvement; The class has worked very hard all quarter; _etc_.) that they were asked to work into a three to five minute script of praise for the class. This praise was to be applied only to the experimental half-classes previous to the combined examination-ratings session. During the second quarter, the same procedure was augmented by also having the teachers provide soft drinks, pretzels and potato chips, as well as praise, to the experimental half-classes during the exam-ratings session. In each quarter, the total CEQ score for the experimental groups was 11% higher than for the control (no praise or no praise/food) group. These differences were statistically significant ($p < .001$) and also administratively significant since the authors (p. 12) reported that this difference represented a movement "from the fifth decile (control group) to the 9th decile (influenced group)". Whether or not these results would occur in other settings or when the ratings and examinations were not offered at the same time remains open to question.

A similar manipulation involves excluding the teacher from the ratings setting to prevent the teacher from employing similar influence tactics. Since it is rarely clear what influence tactics the teacher normally employs, this procedure may be apparently easier to define operationally while still remaining unclear as to the real change that exclusion creates. Cooke (1952) and Flood Page (1974) present data that can be interpreted to mean that the presence of the instructor during the ratings period may produce very slight increases in the ratings scores. Wood (Note 3) included the following sentences in the directions printed at the top of the rating form in order to attempt to standardize procedures, "Some departments will use this data for administrative decisions...this form should be presented at the beginning of a class period--with the teacher absent from the room...a student should return all forms to Room...". The twelfth and last question on the front of the form asked "Was the instructor in the classroom when you completed this form: 1) Yes 2) No". The directions and the question both were intended to cause the teacher to leave the room during the ratings procedure. For reasons still not clear, when the mean class scores for this question were examined only 82 (4%) classes produced mean scores of exactly 1.0, while only 910 classes (44%) produced mean scores of exactly 2.0, and the remainder of the classes apparently disagreed as to whether or not the instructor was or was not in the classroom when the form was completed. When class mean scores for this question were split at the midway point (1.5), 16% (336 classes) of the classes indicated that the instructor was in the room and 84% indicated that the instructor was not in the room. The ratings of lecture skill and clarity were slightly (around 2%, but eta square < .003) but significantly ($p < .01$) higher for those faculty that the majority of the class reported to be in the room, while ratings of rapport or student learning and global ratings of the teacher or the course were not different across these two categories of teachers. Since teachers remaining in the room were violating clearly stated directions that were known to the students, these slight differences may not be characteristic of situations in which the exclusion request is only presented to the teacher and not to the students. The Pasen, et. al. (1978) study combined the presence of the instructor with the use of differently described purposes for the ratings data since the instructor presented the non-administrative use directions while the department chair presented the "administrative uses" directions to similar students in the instructor's absence. The administrative uses--plus instructor absence combination proved much more effective in increasing ratings--especially the rapport ratings.

Another type of influence may occur when students are not only made the target of teacher influence or teacher presence during the ratings situation, but when teacher influence may extend to the post-ratings period because the teacher is able to identify and possibly reward or punish students that produce high or low ratings. Most student raters are permitted to be anonymous, and several systems were designed to withhold even these anonymous results from faculty until faculty have turned in their class grades and thereby lost a major opportunity to reward or punish entire classes because of high or low ratings. Brooks and Levenson (1974) and Colliver (1972) provided students with the option of signing or not signing their ratings and reported that those signing their ratings provided slightly more positive responses. Cook (1952), Pratt and Pratt (1976), Ronning and Walsh (1977), Sharon and Bartlett (1969), Smeltzer and Hartner (1934), Stang and Seiler (1977), and Stone, Spool and Rabinowitz (1977) all produced some evidence supporting the belief that requiring students to give up their anonymity might increase their ratings to a slight degree while Blum (1936) found no such effect. Feldman (in press) has suggested that the effect of anonymity upon ratings may vary from teacher to teacher and may depend upon the students' perception of the vindictiveness of the teacher. Another moderating variable may be the students' perception of the degree to which the teacher has the power--as well as the willingness--to apply rewards or punishment. Neither potential moderator variable was measured in the reported research.

Factors Influencing Student Ratings: Some Conclusions

Factors such as expected grades, class-size, and the students' pre-enrollment attitudes toward the course instructor and/or the course are all likely to influence student ratings. Whether or not these factors are evidences of the invalidity or validity of student ratings remains to be determined. It is probable that expected grade partly reflects the students' (or classes') actual learning in the course while also reflecting attitudes and procedures of the instructor which may be related to both cognitive and affective outcomes. Whether or not the institution or increased use of a student ratings system influences the grading policy of faculty seems to have been discussed but never (rarely?) investigated.

Class size also is likely to interact with the use of different instructional procedures to jointly produce somewhat different levels of ratings for many teachers. Whether or not specific patterns of instructor styles or teaching skills also influence this relationship remains unknown. Whether or not class size influences outcomes other than ratings (e.g., cognitive gain) is now under considerable investigation but the results of the preliminary studies have not been widely reported.

It is possible that, at least for some teachers and for some students, smaller classes permit more effective learning and instruction.

Although it is likely that individual faculty may be able to improve their ratings through various influence techniques or through careful selection of ratings items, much of this instability and potential invalidity may be reduced through the standardization of administrative procedures and the use of a core of common, global ratings items for normative or personnel purposes. The statistical partialling out of the effects of pre-enrollment interest, class size or expected grade has been suggested as a partial reduction of ratings "bias" associated with these factors, but such procedures also may remove some of the legitimate reflection of an instructor's ability--especially when pre-enrollment attitudes are obtained through use of the post-course ratings questionnaire. Although one should probably be very wary of the normative use of data from ratings items asking students to "Grade your instructor (or course, class, experience)," whether or not one would restrict the normative use of global items that reflect "rapport" in favor of items reflecting only "pedagogic skill," is an open question. The use of data from only "skill" items for personnel purposes might reduce the effects of class size and expected grade but might also orient the ratings toward only one type of "ideal" teacher, the clear, enthusiastic lecturer.

Another Approach to the Criterion-Related Validation of Student Ratings

Perhaps the most attractive approach to the concurrent validation of student ratings of teacher effectiveness involves the comparison of ratings with student test scores. However, unless the test scores are themselves reliable and valid reflectors of student accomplishment of all course objectives, and unless the students specifically are asked to rate their own cognitive (or affective?) rather than various characteristics of the teacher or course, such a procedure, even when possible, represents a rather indirect and problematic approach to the concurrent validation of ratings results. A similarly indirect and problematic procedure involves replacing test scores with the criterion of colleague ratings of teacher effectiveness--or with other indicators of faculty success such as research productivity or effectiveness in the performance of faculty service activities. Although this approach has the added attraction of not requiring different faculty to teach toward the same objectives, it often depends upon the reliability and validity of peer or administrator judgments concerning teaching performances which the peer-raters did not view or experience--or upon other criteria which are themselves difficult to evaluate and/or possibly unrelated to instructional effectiveness.

Student Ratings Versus Peer Ratings

The reliability of the ratings of individual students tends to be relatively low (Gillmore, 1973) and is usually increased by the use of fifteen or more student raters and/or the use of multi-item ratings scales. Unless a similar number of faculty can be exposed to a teacher's performance, in person or via videotape, there is little reason to believe that the reliability of the ratings of peer observers is likely to be high enough to justify the use of these ratings as a correlate to student ratings. Love, et al. (1977) and Love and Sandoval (1978) produced reasonable observer reliabilities with small numbers of observers by training their observers, but even then, they faced problems with reliability and with the interpretation of their observer scales. Centra (1975) employed peer ratings scales which were more comparable with his student rating scales, but his peer raters were so generous in their ratings of their colleagues that the results were useless as correlates for student ratings. Some of the outcomes of the Centra (1975), Love and Sandoval (1978) and Love, et al. (1977) studies also lead the reader to question if even the reliable ratings of faculty colleagues are truly comparable to the ratings of students. It is very possible that a student and a visiting faculty member will have different perceptions of a teacher's performance due to the difference in exposure and due to the difference in rater knowledge and in perceived evaluation-press. To fairly judge a colleague's teaching effectiveness, faculty members might actually have to visit a series of classes and possibly take one or more course examinations. Since faculty appear to be unwilling to visit one anothers classes (Cross, 1977), the common employment of such a series of extensive visitations by many faculty appears unlikely.

The more common approach to the comparisons of student ratings and faculty judgments involves asking the faculty to produce general judgments of their colleagues' teaching effectiveness--usually with little or no reason to believe that the faculty raters have ever visited the classes of those colleagues that they are rating. In most cases, the faculty raters have not even been provided with the goals, examinations, syllabi, teaching materials or other teacher-created or teacher-chosen artifacts employed in instruction. Presumably, these peer judgments are based upon comments by those students that they know or have overheard, observation of faculty performance in non-classroom settings, conversations with their colleagues, and similar evidence. Guthrie (1949, 1954) has reported that _ad hoc_ faculty committees of five to seven persons produced judgments of their colleagues that correlated well (_r_ = .64 to .76) with judgments of different sets of five to seven colleagues, and also correlated moderately with student ratings of teaching effectiveness (_r_ = .48). Swanson and Sisson (1971) reported that _the same_ 12 faculty colleagues produced similar judgments of their departmental colleagues' total (teaching, research and service) effectiveness (_r_ = .72) three months later, and that these overall peer judgements correlated to some degree (_rho_ = .70 and .28) with two sets of student ratings of the same faculty. Wood (1977, 1978) reported that while departmental peer ratings or rankings of their colleagues' teaching effectiveness were relatively stable over a four year period (_r_ = .43 to .82), these peer judgments were generally unrelated to student ratings of teaching effectiveness. However, when peer ratings were compared with student ratings, the correlations were in the low .60's, but when a ranking or nomination system replaced a peer ratings system, the peer-student correlation dropped to the .20's and low .30's. The level of peer-student ratings correlations reported by other researchers ranged from .11 to .49 by Braunstein and Benston (1973) and .27 and .28 by Aleamoni and Yimer (1973), to .48 (Guthrie, 1949), .49 (Costin, 1966), .62 (Blackburn & Clark, 1971), and .69 (Maslow & Zimmerman, 1956), to a high of .84 by Ballard, Reardon and Nelson (1976). The results of the previously mentioned Morsh, Burgess and Smith (1956) study in a military setting included a high correlation between supervisor and peer-instructor rankings (.77), moderate correlations between student ratings and learning gains (.41 to .46), but negligible correlations between supervisor or peer-instructor ranking and student learning gains. Unfortunately, the majority of the other reported investigations of the relationships between faculty and student ratings of teaching effectiveness failed to include any similar, additional indicators of teaching effectiveness.

Research Activity Versus Student Ratings

With a few exceptions, most researchers from Guthrie on have produced data to support the position that research and teaching tend to be relatively independent faculty functions. Faia (1976) and Hoyt and Spangler (1976) have reviewed nearly a score of studies in which indicators of research productivity have been correlated with indicators of effective teaching. In general: (1) indicators of teaching effectiveness (usually student ratings) are uncorrelated with indicators of research or publication; (2) certain hiring patterns, reward systems and academic disciplines may be associated with moderately positive or moderately negative correlations between teaching and research effectiveness; and (3) faculty ratings of teaching effectiveness are more likely than student ratings of teaching effectiveness to be correlated with indicators of research productivity. This potential bias in peer judgements of teaching effectiveness also emerged in the Hildebrand, Wilson and Dienst (1971) finding that faculty, but not students, tended to include research and service descriptors in their definitions of the "best" versus the "worst" teachers. The reverse bias toward research productivity may be exhibited by students who have rated faculty lack of knowledge as a relatively minor teacher problem in comparison with faculty inability to "stimulate interest of students in subject matter," "relate subject matter more to student interests," and in general to present their knowledge in an interesting manner which permits students to reconstruct this new knowledge so as to permit students to connect subject matter with the students' past experience and present interests (Morstain & Gaff, 1977).

Summary

Although peer judgments of teaching effectiveness are frequently correlated with student ratings, peer judgments appear to be at least as open to bias and error as are student ratings. Peer judgments might more appropriately be focused upon research and service activities and upon those aspects of teaching that are more easily exposed to faculty view--teacher knowledge of the academic discipline, the adequacy of examinations and examination procedures, the adequacy of syllabi or course hand-outs or audio-visual materials, and the meeting of professional obligations such as being available during posted office hours.

Research effectiveness and student ratings are usually not correlated, but this relationship may be influenced by institutional or discipline-related hiring and reward procedures. Research effectiveness is somewhat more likely to be correlated positively with faculty ratings of teaching effectiveness than with student ratings of faculty effectiveness. Students, unlike faculty, also may be more interested in their teachers' ability to stimulate interest and translate subject matter than in their teachers' actual depth of knowledge or academic credentials.

Utility: Do Student Ratings Help?

The Improvement of Classroom Instruction

Faculty Attitudes

Student ratings are usually the responses of students to questions which ask them to describe or evaluate aspects of a particular set of instructional interactions that take place in a college classroom and that involve students and their instructor(s). Centra's (1976) survey of postsecondary faculty development practices indicated that student ratings are the most commonly employed technique to measure the effectiveness of instruction, and that student feedback was judged to be moderately effective in assisting teachers to improve. Guion, Hutchinson, Klein, Stutz and Wood (1977) surveyed the faculty at a medium sized Midwestern state university (15,000+ students) and reported that: 91% of the responding faculty agreed that student ratings should be used for teacher self-improvement; 79% agreed that ratings should be used to assist the chairperson (or others) to assist teachers; 68% agreed that ratings assisted them to assess and improve their teaching; but only 38% agreed that ratings should be used to inform students about teacher/course characteristics and only 59% agreed that ratings should be used as a necessary component of personnel decisions. Ratings appeared to be perceived by the majority of these faculty to be best employed to assist teachers to improve their instruction (and moderately useful in performing this function) and least appropriately employed to provide students with information concerning courses or teachers. Wood (Note 3) included two questions that teachers were asked to complete when using his student ratings questionnaire: (1) To what degree are your students capable of assessing the quality and value of their experience in this class? (very capable 1 2 .... 6 7 not capable); and (2) How useful do you find the type of data produced by this type of form? (useful 1 2 .... 6 7 useless) Only 9 to 10 percent of the responses were at the extreme positive side of the scale, 24 to 25% were at the next most positive point, and 58 to 62% were on the positive side of the middle point of the scale. Evidently most of these users of a ratings scale were at least moderately confident of their students abilities to rate the quality of their class experience and moderately hopeful about making some use of the resulting data. A survey of faculty at the University of Nebraska (Cross, Mazur-Hart, & Kess, 1976) reported that only 10% of the respondents indicated that teaching was less satisfying than other aspects of their work, 66% regularly collected student evaluations for their own use (8% collected them occasionally for their own use, and only 18% reported that they collected them when required to do so or when needed for the personnel file). Approximately 80% of these faculty claimed that informal recognition from peers or students was very important or quite important to them as a motivator for improving their teaching. Although more of these Nebraska faculty also indicated that scholarly productivity, learning more about teaching techniques, and especially, updating course material, were more important to the improvement of their performance as a teacher, at least 40% also indicated that student evaluations were extremely or quite important and another 30% claimed them to be moderately important.

These survey results can be interpreted as meaning that at least some college faculty, and perhaps a majority of them, are willing and able to employ the data produced by student ratings to improve their classroom performance.

Experimental Research

There are several experimental studies that have demonstrated that pre-college teachers or administrators have made improvements subsequent to receiving ratings data. Tuckman and Oliver (1968) obtained student and supervisor (principal, vice-principal) ratings of 286 high school teachers of vocational subjects in the late fall, randomly assigned the teachers to four feedback conditions, and provided teachers with: (1) no feedback; (2) ten graphs showing the teachers' position on the ten student ratings scales and also a summary of student commentary;

(3) similar feedback from the supervisor; or (4) feedback for both students and supervisor. Twelve weeks later, student ratings were again collected to identify change scores, and 15 additional teachers were added as a post-test-only control group. Although the spring ratings tended to be lower than the fall ratings, those teachers in the student feedback group were significantly better (less low) than those in the no-feedback group. Unexpectedly, and for unknown reasons, supervisor feedback appeared to produce a greater negative shift than no feedback. Bryan (1963) produced a similar effect from student feedback (to secondary teachers) in a smaller and less complex study, and Gage, Runkel and Chatterjee (1963) similarly reported a positive shift of subsequent ratings that was attributable to providing middle school teachers with student descriptions of the ideal teacher.

None of these experimentally controlled, feedback-associated changes in subsequent ratings were very large, but each study did provide evidence that student ratings data or student descriptions of ideal teachers could assist pre-college teachers to change their instructional practices enough to be credited with a positive (or, at least, a less negative) change in the opinions of their students.

Perhaps the largest and most complex pre-college feedback study was conducted by Daw and Gage (1967) when they experimentally provided or withheld teacher feedback from 455 elementary school principals and other teacher supervisors. Although Daw and Gage (1967) were successful in demonstrating that those principals who received teacher feedback later elicited higher ratings from their teachers, the experiments could find no relationship between indicators of their principals' experience and their tendency to employ feedback to improve their subsequent ratings.

Evidently, in the precollege setting, providing teachers or principals with the ratings or other feedback produced by those students or teachers they supervise is enough to cause the teachers or principals to change so as to be rated more similar to their supervisees' ideal. A similarly limited number of studies of college teachers has not produced similarly consistent findings. Centra (1972), Emmer, McBurnette and Davis (1974), Miller (1971) Pambookian (1974) and Wood (Note 6) found that college teachers do not change when provided with the results derived from student ratings. However Braunstein, Klein and Rachla (1973) and Marsh, Fleiner and Thomas (1975) have found that the provision of ratings data does improve subsequent ratings, while Aleamoni (1973, 1974), Erickson and Erickson (in press), Erickson and Sheehan (1976) and Hoyt and Howard (1977, 1978) have demonstrated that the use of ratings data as part of a faculty improvement system can produce increases in subsequent ratings--at least with faculty who volunteer to be assisted.

Miller (1971) was one of the few researchers to use both student ratings and students' examination scores as criteria when he examined the effect of mid-semester ratings feedback upon the end-semester ratings and exam scores of thirty-six, graduate assistant, "discussion leaders" attached to three college courses. Discussion leaders randomly assigned to the feedback condition didn't elicit ratings that differed from their control colleagues, but may have produced higher test scores in one of the three classes. The leader's ratings of the value of the student responses to different ratings items had no identified effect. Centra (1972) overcame the problems associated with Miller's focus upon a limited number of graduate assistants and added the post-test-only control group employed in previous pre-college research when he provided

177 college faculty from five colleges with mid-semester ratings results which permitted them to compare their results with those of other teachers. Centra withheld similar results from a control group of 188 college teachers and obtained end of semester ratings from 137 of the 177 experimentals, 159 of the 188 controls, and from 140 post-test-only teachers. Although the teachers' subject area, sex, experience and college affilation influenced the ratings, the experimental treatment didn't. However, teachers who rated themselves more favorably than they were rated by their students may have improved their ratings on several items. When 8 teachers from the feedback group were compared with 13 from the control group and 21 other teachers at the end of the subsequent semester, the now-doubly rated experimentals produced some higher item scores than the singly-rated controls (who had never been given comparative data) or the never-rated twenty-one new controls. Emmers, McBurnette and Davis (1974) and Pambookian (1974) employed smaller samples of college teachers and produced results similar to those produced by Centra's (1972) main study--no overall improvement for those who received ratings feedback but some reason to believe that discrepancy between student ratings and teacher self-ratings might be associated with an improvement on some item scores. Braunstein, Klein and Pachla (1973) and Marsh, Fleiner and Thomas (1975) similarly provided mid-term ratings data to some teachers but not to others. Braunstein, et al., (1973) found that their feedback group produced superior end-term ratings. Marsh, et al., (1975) found that their feedback group elicited significantly higher ratings scores on 8 of 46 items and on three of eleven factors while not eliciting higher test scores. An interesting outcome indicated in Table 1 (p. 837) of the Marsh, et al., (1975) study was that the

"instructor approachability" factor and the "readings" factor were the two, significantly improved, multi-item ratings factors, but ranked next to last ($\underline{r} = .08$) and fourth from last ($\underline{r} = .21$) in terms of correlation with final exam scores while the best predictor of final examination scores ("course organization," $\underline{r} = .55$, $\underline{p} < .01$) was last in terms of apparent improvement. However, the numbers of faculty involved (unreported, but probably fewer than 9), and their limited control over their multisection course (which was directed by a course director who provided the course outline) do not permit more than speculation as to the implications of this data. Wood (Note 6) provided the use of a student ratings scale to graduate assistants enrolled in a "professional development" workshop that preceded the beginning of their first of three successive academic quarters of teaching. One-hundred and thirteen teaching GA's used the scale with their students at the end of the first academic quarter, one hundred and twenty (59 of whom had used the scale the first quarter) used it the second quarter, and 91 (50 repeaters from term 1, 56 from term 2) used it the third quarter. Neither one nor two uses of this scale appeared to differentiate previous users from first-time users, and subsequent ratings were not related to the number of terms of previous teaching experience. Possibly the typical graduate teacher's lack of course control and the increasing pressures of graduate life functioned to reduce increases in subsequent ratings attributable to previous experience with teaching the same course with or without ratings feedback.

Ratings As One Part of a Faculty Improvement System

Centra (1976, 1977a) has documented the recent expansion of faculty development centers at postsecondary institutions and, with others, (e.g. Bergquist & Phillips, 1975; Crow, Milton, Moomaw & O'Connell, 1976; Gaff, 1975), has provided partial descriptions of this phenomenon. Centra (1977a) concluded that "most of these units had existed only two or three years and had not yet been evaluated adequately" (pp. 53-54), and Hoyt and Howard (1977) reviewed several attempts to evaluate the effect of these centers and concluded, "the literature is extremely sparse and the studies reported are uncommonly simplistic...apparently, participants in faculty development programs have generally expressed satisfaction with them...there is some evidence that teaching methods may change in directions considered desirable by teaching authorities...no dependable evidence regarding impact on students was reported" (p. 2).

Three of these faculty development (or instructional development) centers have reported results achieved by the provision of ratings data that was accomplished by additional supportive and focusing activities.

Lawrence Aleamoni was instrumental in the development and description of the Illinois Course Evaluation Questionnaire (The CEQ), (Aleamoni, 1972; Spencer & Aleamoni, 1969), the package of computer programs (MERMAC by Bussell, Costa, Spencer & Aleamoni, 1971) used by many institutions to analyze and present questionnaire and test data, and the development of instructional improvement services at the University of Illinois and at the University of Arizona. He has described (Aleamoni, 1973, 1974) a relatively simple instructional development procedure which involved: (1) the use of the CEQ with one or more classes; (2) a "15 to 20" minute conference with Aleamoni during which low subscale scores, low item scores and/or the instructor's recall are employed to jointly explore and identify specific instructional problems and possible remediation strategies; and (3) a subsequent use of the CEQ with a similar class to identify change in ratings scores. If the source of difficulty is not identified, additional procedures (such as the use of other ratings or peer observations) are employed. Aleamoni (1973) has reported that the decile increase in subsequent terms with similar courses (on the lowest subscale decile) was 3.94 for 17 teachers who participated in an individual conference and .57 for the seven teachers "who were not able to participate in the individual conferences" and who were "available to use as a control group" (p. 47). Although the numbers of teachers were small, the statistical test nonexistent, the control of questionable similarity to the experimental group, and the class sizes unusually small, the results are uniformly positive.

Erickson and Sheehan (1976) and Erickson and Erickson (in press) were instrumental in the development of a similar but perhaps more complex ratings and consultation process than was originated at the University of Massachusetts by Melnik, described by Melnik and Sheehan (1976), and characterized by Bergquist and Phillips (1977) as "perhaps the most powerful methodology yet conceived for the actual improvement of in-class teaching" (p. 78). The process has been adopted by "dozens of colleges and universities...as a component in their faculty development programs and hundreds of faculty have participated in the process" (Erickson & Erickson, in press, p. 5). The Erickson and Sheehan (1976) study reported that 40 faculty (volunteers) were randomly assigned to one of three levels of treatment (1=collection of ratings and other data describing the teacher, but no feedback; 2=collection of data and a consultation concerning it; 3=data collection, consultation concerning data, and continuing consultation with continuous data collection and feedback) and that the most complete treatment (#3) elicited a positive rating from involved faculty. In general, more treatment seemed to produce higher end-term student ratings--although the results were mixed and not statistically compelling. The authors also reported that it was very difficult (perhaps impossible and possibly unethical) to withhold aspects of the process from those faculty that wanted more than what their random assignement to level of experimental treatment entitled them. The problems involved in withholding treatment (or aspects of treatment) from clients has been documented in the medical literature and was also presented clearly in an issue of Science (18 November 1977) devoted to this conflict between the desire of the confident, client-focused, clinician to try everything and anything that might help the client and the need of the researcher (e.g. Tukey, 1977, 679-684) to withhold the presentation of some untested procedures in order to determine the impact of one or more relatively untested procedures upon clients with different characteristics.

Erickson and Erickson (1979) reported a study that was a simplified replication of the Erickson and Sheehan research. Erickson and Erickson (1979) randomly assigned 31 volunteer college faculty to either a full treatment group (early semester and late semester student ratings plus feedback of the early ratings results and full, continuous consultation and data gathering) or to a control group (only early and late semester ratings, no consultation and no feedback until end of semester). The last semester student ratings and the late semester teacher self-ratings were higher for the experimental group ($p < .05$ to .01; omega square = .13 to .30) for each of the three ratings factors, the experimental group improved more on the ratings reflecting skills that had been targets of the consultation/improvement efforts than on other ratings. The experimental faculty also rated the process positively and as worthwhile. The second Erickson and Erickson (in press) study involved the identification of instructors who had experienced the ratings-consultation-ratings process and who were again teaching a similar course. These 20 (of the 36 college teachers who had experienced the full consultation process, 14 were not teaching comparable courses and two others didn't agree to participate) teachers were asked to provide early semester ratings from their current class, and these ratings were compared with their pre-consultation, early semester ratings that had been elicited one (2 teachers), two (16 teachers), three (1 teacher) or four (1 teacher) semesters earlier. Three of the twenty teachers produced lower ratings (not significantly lower), seven more produced (non-significantly) higher ratings, and the remaining 10 produced significantly higher ratings. The mean increase for the 17 improved (or still improved) teachers was nearly half a scale point on a four point scale with a standard deviation somewhere around .68. The three teachers who dropped in ratings lost a mean of .16 points. No data was provided concerning the ratings of these teachers immediately after the consultation process that typically occurred two semesters earlier—or about their end term ratings in the examined semester.

Hoyt and Howard (1977, 1978) reported the description and results of two studies similar to the two Erickson-Erickson studies as well as the results of a series of faculty surveys focused upon faculty opinions of the Kansas State IDEA system. The experimental study was conducted by Howard at Wichita State and involved: (1) inviting all faculty to apply for participation in a new faculty development program; (2) randomly selecting 16 of the 68 applicants for the experimental group and 16 others for the control group; (3) asking the 32 faculty to administer both a mid-semester and an end-semester student ratings questionnaire; and (4) enrolling the 16 experimental faculty in two, eight-faculty groups which ran from eight to ten weeks and included discussion, class observation and feedback, and videotaping and analysis of teaching performance. The experimental group was significantly (p < .01) superior to the control group in terms of total ratings (final ratings controlled for midterm ratings via an analysis of covariance), overall ratings as a teacher (one item) and on three of the remaining 12 ratings items. The control group was slightly but not significantly superior on only one item "made it easy to take lecture notes" (F = .015). The superiority of the experimental group was generally small in magnitude (the mean control group total score decreased 2.3% from midterm to final ratings while the experimental group mean total score increased 1.3%). The non-experimental investigations involved: (1) comparing the scores of 263 faculty who had produced ratings of the same course during two different semesters in the 1969 to 1972 period; (2) comparing the ratings of 348 faculty who had reported ratings from the same course during different semesters of the 1973 to 1975 period; (3) categorizing the latter 348 faculty as to interaction with the program director (several interviews or enrollment in a course on college teaching; some direct contact via an interview or group session; and no contact); and (4) comparing the earlier ratings for these teachers with later ratings in a search for improvement. The second set of comparisons were examined for increases attributable to contact with the director as well as attributable to the use of ratings. Significant improvement occurred over time for both sets of teachers although the overall gains were not great (around 2 to

2½%--or around 17% to 20% of a standard deviation), and the degree of contact with the project director was positively related to the degree of improvement. Since faculty motivation, a possible general increase in student leniency, possible changes in student populations, possible effects due to grade inflation, and a host of other alternate explanations for ratings increase were uncontrolled and unexamined, this modest improvement in ratings by users of the system may or may not be attributable to the system--and many not be generalizable to non-volunteer faculty. Hoyt and Howard (1977) also provided the responses of system users to several questionnaires, and these responses--and the reported increase in numbers of faculty volunteering to use aspects of the system--also provide evidence that the IDEA faculty development system is meeting its goals. The various progress reports by Cashin and Slawson (1977a, 1977b) and Hoyt and Cashin (1977) also provided evidence of the extensive and growing national use of aspects of this ratings and feedback system.

Evaluation for Improvement: Paradoxes

There are several paradoxes associated with the attempt to improve faculty performance through the use of student assessments of this performance. Perhaps the most obvious is presented by the majority of the faculty or instructional development centers that are, at the same time, (1) encouraging teachers to make greater use of assessment procedures because assessment is believed to be of general and specific benefit to the teacher, (2) while these same centers are quite obviously avoiding the organized assessment (for purposes of improvement) of their own operations, processes and impacts--many of which are clearer and more open to assessment than are those of their clients. All the while many college teachers, who may be less well trained in assessment and instructional procedures than any other group of American teachers, regularly assess and publicly report upon the apparent learning of their students, frequently with little regard to even the reliability of their measures, while claiming that student ratings scales with demonstrated reliabilty (and considerable validity for several purposes) do not have sufficient validity for use with their classes and if used must produce results that are not reported to others. The third paradox was presented by Glass (1975) when he stated that:

> The "excellent" state of affairs seems to emerge most surely when those on whom it depends feel that they are not being judged and evaluated, feel that their shortcomings are understood and accepted, feel that their worth--as teachers, as administrators, as professionals--is unconditional, that is, not contingent on how they score on anyone else's scales of merit. And yet it appears that many people move truer and more certainly toward excellence to the extent that they clarify their purposes, measure the impact of their action, judge it, and move on--in a few words, evaluate their progress (p. 9).

Glass provided a partial explanation for the possible harm derived from the imposition of external evaluation by describing some of the results of evaluation anxiety or "the crushing pressure of judgmental averments" and "the fear which (evaluation)...excites in the person evaluated...fear of really knowing what we are doing, why and with what consequences--fear of the feedback earnestly proffered by the evaluator...fear (that) selectively censors the reality experience...confounds thought and action...makes people ignorant, refractory and less than excellent" (p. 10). Although this analogy between the neurotic and extremely defensive client in the office of the cautious and adroit therapist (who "attempts to manage that fear so that the reality messages can be received, than haltingly acted upon, then eventually lived within," p. 10) and the teacher being provided with assessment and consultation from students and/or development staff many be overdramatic, it many not be too far in error. Orlandi and Consalvo (1977) claimed that "most educators who want evaluation at all, want it for someone else...nobody likes the idea of someone collecting evidence that may be used against him/her, no matter how much this use of evaluation is played down, and emphasis placed upon the more laudatory role of formative evaluation in professional development." "Educators also resent the 'interference' in their routine that they associate with evaluation efforts." "Nobody likes to be found at fault for what is wrong, and nobody really wants to change his/her ways" (p. 7).

The Glass paradox was not resolved completely by Glass (at least in his article), and with increasing pressures for accountability in higher education as well as likely pressures arising from an oversupply of college teachers and an undersupply of college students, may present an increasing problem for those involved with teacher assessment and improvement. The other conflicts or paradoxes are characteristic of any procedures (such as the interrelated processes of self-confrontation and external evaluation) which are capable of bringing pain and/or growth--depending upon the intensity of the procedure, the security and style of the recipient, the empathy and/or skill of the transmitting agent (consultant, evaluator, therapist), and the degree to which the information or feedback provides clear and acceptable avenues for change. An extensive body of literature has been created in attempts to identify the characteristics of the information, the recipient, the treatment, and the transmitting agent. Glass' (1975, p. 10) description of the interaction of the psycho-therapist and the fearful but hopeful client is an outgrowth of at least one branch of this literature. Tuchman (1973, pp. 91-131) has provided a similar interpretation of this process as it applies to the provision of ratings-derived feedback to teachers and has compressed his conclusions into "12 rules of effective feedback" (Figure 7, p. 123). Fuller and Manning (1973) have provided an extensive and thorough review of a related self-confrontation-for-improvement device, video-feedback, that has been closely associated with the use of student ratings--in that ratings are often used whenever videofeedback is used with teachers, and videofeedback is an integral component of the widely employed consultation process described by Mednik and Sheehan (1976). Wergin (1977) has suggested criteria for evaluating the effectiveness of developmental centers and their activities, and many of these suggestions involve a combination of the suggestion that centers become more aware of the needs and positions (a kind of evaluator's version of institutional empathy?) of constituents and potential users and an implied but continuous warning that inappropriate self-evaluation or inattention to the interests of other powerblocks in the environment may result in the destruction of the evaluation/improvement center. The closing of the highly reputed instructional improvement center which pioneered the use of the consultation procedure described by Mednik and Sheehan (1976) and the apparent difficulties faced by other faculty improvement projects provide proof that fear of administrative decision, with or without assessment data, can be a real concern and not always a neurotic misperception of the environment.

Do Ratings Help or Harm?

There is a limited amount of evidence--all produced by persons or centers that are paid to assist in the development and improvement of faculty and instruction, and all a product of research strategies which are far more casual and far less informative than procedures that are available in the demonstrated repertoire of the originating institutions, that supports the position that the voluntary use of student ratings may increase the subsequent ratings of college teachers to some degree (probably by a few percent). Voluntary participation in more extensive improvement programs may improve overall ratings as well as shift the focus of instruction (e.g., Howard's two, eight teacher seminar groups apparently caused these teachers to move more toward the discussion approach since ratings related to this approach were those ratings that gave them the most comparative, although slight, improvement). Volunta participation in an intensive counseling and video-tape system may improve subsequent ratings to a greater degree. Since the counseling literature suggests that: (1) two out of three counselors are probably not helpful and may even be noxious (Truax & Mitchell, 1971); and (2) the counselor may be more important than the type of feedback source (Searle, 1969); the selection of counselors and voluntary involvement or choice of counselor may both be very important factors. Fifteen or twenty minutes with Aleamoni or time spent with Hoyt, the Ericksons or Sheehan may assist faculty to improve while consultation time with other people may not help. Fuller and Manning (1973) employed an old saying to make this point, "wolves hire out very cheaply as shepherds" (p. 499). (What they do not say is that even good shepherds are so underpaid that hiring price may not be the best selection criteria.)

Research reported by Fuller and Manning (1973) indicated that self-confrontation can easily produce stress and even disruption of function; Tuckman (1973) suggested that "reception of feedback must not involve more than low risk....models for change and support for change must be provided" (p. 123); and Fuller and Manning (1973) concluded that "those who are vulnerable and without capacity to change can be damaged....(but that) self-confrontation now seems to us more promising than we had dared hope and more dangerous than we knew to fear" (pp. 511-512). Since the college teacher population more resembles those persons more likely to be helped by self-confrontation (the YAVIS population; Young?, Attractive?, Verbal, Intelligent, and Successful) rather than those likely to be most harmed (the HOUNDS; Homely, Old, Unattractive, Nonverbal and Dumb), mild self-confrontation devices like ratings feedback (even without much focus or clear suggestion for alternate but reasonable action or adequate support) should help more than harm. Needless to say, there has been no major research effort to determine if the introduction or employment of ratings systems improve or negatively influence such important factors as faculty morale, faculty satisfaction, student satisfaction, or faculty-student cooperation.

The Use of Ratings Data to Make Better Decisions

Although the data derived from student descriptions of teaching may be employed to suggest avenues of improvement for subsequent teaching, and may even lead to the improved learning and/or satisfaction of students, researchers such as Gillmore (1972) and McKeachie (1969) have suggested that data from student ratings also may be employed: (1) to assist personnel committees to make better decisions; (2) to assist students to make better choices of classes or instructors; (3) to assist departments or colleges to improve their programs; and (4) even to sensitize students and faculty to important aspects of educational processes and products. These additional uses of ratings data have not been investigated and reported upon to any major degree.

Student Ratings and Personnel Decisions

Guthrie (1949) noted that tenured faculty at the University of Washington tended to receive superior student ratings and suggested that this might provide some reason to believe that his University's tenure-decision procedures (which included the use of data drawn from student and faculty ratings of teaching performance) were effective and were responsive to student opinions. Downie (1952) reported higher student ratings for full professors, and Gage (1961) reported higher ratings for both full and associate professors, and these findings might be taken as support for tenure and promotion systems that are sensitive to student ratings--although a number of alternate explanations are also possible and reasonable. However, Aleamoni and Yimer (1973) found no ratings difference across faculty ranks, and Frey (1978) reported that while more experienced faculty received higher ratings on items reflecting pedagogic skill, faculty of lower rank and less experience tended to be rated more highly on items reflecting rapport. Centra and Creech's (1976) survey of over 8,400 college faculty indicated that only teaching assistants (N=243) differed in their global effectiveness ratings (significantly lower) from other faculty ranks (which didn't differ from one another), but that global effectiveness ratings tended to increase with years of teaching experience--with a slight reduction in a teacher's later years.

Since personnel decisions, years of experience, individual faculty decisions to stay in or leave teaching, changes in teaching assignments and changes in faculty control over their environment all occur simultaneously over time, and since all may influence student ratings, it is very difficult to untangle the various relationships to attribute likely causation. When colleges have changed their hiring policies to focus more upon the research potential (or previous teaching experience) of entering faculty, these changes are also likely to influence any attempts to compare pre-change and post-change faculty in terms of student ratings.

Other Evidence That Ratings are Employed as Part of Personnel Decisions

The use of student ratings in the determination of personnel decisions may be mandated or suggested in faculty charters or administrative procedures, but there is little evidence that indicates whether or not such information is actually employed or to what degree student-produced data actually influences decisions. Rood's (1977) analysis of the legal issues in faculty termination led Rood to categorize termination cases as: (1) being "on the basis of financial exigency" (p. 142); (2) involving failure to grant tenure; or (3) involving questions of "competence, performance and character" (p. 142). Court support of decisions to terminate because of financial exigency appeared to hinge much more upon the institution's ability and willingness to provide faculty with due process and also to have available a clear set of objective criteria upon which to base termination. According to Rood, the courts also seemed to consider that non-tenured teachers clearly held a probationary status and that the typical non-tenured teacher has "no property interest in future contracts and is not guaranteed by the due process clause either a statement of reasons for his termination or the opportunity to have a hearing" (p. 132). Even the release of a tenured teacher for reasons of competence or conduct seemed to hinge more upon the demonstrated impartiality of the hearing board than upon the use of objective criteria of competence. Quite obviously, the use of student opinion could easily be excluded from consideration of faculty termination--even though many of the reasons not to rehire employed in the Raney case (Rood, 1977, pp. 147-148) were types of information likely to have been reported by students (unsuitable grading policy, "tough" attitude toward students which caused excessive dropouts, severity of grading, sarcasm toward students, general reputation among students, faculty...as a contentious group). Alm, Ehrle, and Webster (1977) reported upon the procedures employed by Mankato State University to reduce its faculty by 150 positions over a five year period. These authors described a set of clear, objective criteria ((1) release non-tenured faculty first; (2) identify programs with low student to faculty ratios; and (3) release tenured faculty in low ratio programs in reverse order of seniority) which did not employ either student ratings or any other indicators of faculty quality. These authors even suggested that such a financial exigency situation is "no place to deal with...trimming out so-called 'deadwood', or otherwise influencing the composition of your institution" (p. 161). Although student ratings may be employed in procedures involving faculty dismissal, there is little published evidence suggesting that they are used--or even should be used.

Whether or not student opinions are employed in more positively oriented personnel decisions is even more difficult to determined due to the global nature of the decision to be made, due to the multiplicity of possible criteria of faculty effectiveness, and due to the failure of most institutions to document and report their actual procedures. Roose (1977) has reported the use of Social Judgment Theory (Hammond & Adelman, 1976; Hammond, Stewart, Brehmer & Steinman, 1975) and an associated methodology in order to attempt to determine the actual weights assigned to different variables by faculty who were asked to make judgements concerning the salary and rank of other faculty. Unfortunately, the inclusion of student ratings of teaching performance was precluded because of a variety of institutional and individual decisions that occurred during the early stages of the investigation. In general, the faculty "judges' assignment of rank and salary were highly predictable by equations which included only a small subset of the 28 variables" used to describe faculty, (Roose, 1977, p. 63) although different faculty weighted the descriptive variables differently, and every descriptive variable was employed to some significant degree by at least one of the judges. The highest degree earned by a faculty member was the most important predictor of both rank and salary, with years at the institution, years since the highest degree, and publication of books and articles also functioning as good predictors. Variables such as percentage of time spent teaching, teaching awards, and self-ratings of teaching evaluations all ranked quite low as predictors of rank and salary--with only variables relating to supervision of student research and "professional creativity" ranking lower. Although the substitution of student ratings for faculty self-reports of teaching activity might raise the weighting of teaching activity in subsequent investigations of faculty or administrative personnel judgements, such a shift is problematic at best.

Use of Student Opinions in Personnel Decision-making: Conclusions

There appears to be little evidence that data drawn from student ratings-questionnaires is weighted very heavily in the decision-making processes of personnel committees at the college level--although the available evidence is very limited. Social Judgement Theory and its associated investigative procedures appear to offer a potentially useful approach to the empirical explication of the murky arena of personnel decision-making at the post-secondary level.

Student Ratings and Program Evaluation

Braskamp (1978), Clark (1977), and Harnett and Centra (1976) have all described approaches involving the use of student questionnaires to gather data to be used for the improvement of college departments or programs. Braskamp (1978) reported: that student judgments possessed considerable discriminant and criterion-related validity; that student judgments verified or documented faculty judgments in a majority of departments; that graduate students were more likely than undergraduates to value departments with older and more experienced faculty; and that student satisfaction was uncorrelated with the scholarly publication of departmental faculty. Harnett and Centra (1976) failed to find a set of characteristics that clearly differentiated departments as to their apparent effects upon student learning but claimed to be able to identify departments that did produce different effects upon their students. Clark (1977) has reported that a majority of department chairpersons claim to value student judgment as to teaching quality or even concerning the learning environment even though other effectiveness criteria are more important for use with university or beyond-university audiences.

There is little or no published evidence that student opinions are actually employed in decisions regarding departmental modifications, but there is little or no evidence to the contrary either.

Ratings and Student Choice of Classes.

Although the publication of the results of student-ratings questionnaires exists at many colleges, there is little or no published evidence attesting to the effect of such publications upon student decisions or upon faculty function or attitude. Guion, et al. (1977) reported that the faculty of at least one university were generally opposed to the open publication of student opinions of faculty effectiveness.

UTILITY: Why don't ratings work?

Two of the major functions of student ratings procedures were described by Hildebrand, Wilson and Dienst (1971: "One major aim of this study is to define and describe effective teaching so that instructors can be helped to improve....The other major aim is to find more valid, reliable, and effective means of incorporating the evaluation of teaching into advancement procedures...; the incentive thereby provided will encourage instructors to devote the study, time, and effort necessary to do their best, and the status of teaching will increase" (p. 1). However, there is little published evidence that indicates that teaching is directly improved through use of student ratings procedures or that the development and employment of ratings procedures has shifted the focus of personnel committees or administrators from research toward teaching. Although it is possible that the development and employment of ratings scales have improved teaching performance or increased the status of teaching without such changes being extensively documented and reported, it also is possible that most ratings systems and the situations in which they are employed tend to inhibit their effective functioning.

Problem 1: Unclear Objectives

Student ratings procedures are similar to the use of general achievement surveys in the public schools in that their wide availability, frequent usage, and the multiple purposes of each system tend to permit an extensive use by people or institutions that have NOT: (1) clearly determined the specific purpose(s) of the evaluation procedure; and (2) carefully chosen the instrument(s) most likely to produce data to accomplish this(these) purposes(s). The questionnaire or ratings technique is especially sensitive to this problem of thoughtless usage because the scales appear to be so easy to construct. The wide availability of student ratings items and scales and the availability of computerized scoring and analysis systems further compound this problem by permitting groups of interested students and/or faculty to construct their own systems with such little temporal or financial cost that they don't even have to explain their actions to those who hold the pursestrings--much less to those expected to use the resultant information. The effective employment of student ratings data is similar to the effective employment of the results of achievement surveys in that unless all users are informed about and guided concerning the interpretation and uses of the data, there is a strong tendency to ignore the confusing mass of apparently meaningless numbers and put aside the data until some later and more convenient time (which may never come).

Feedback to Improve Classroom Instruction

Specificity of Information

The multiple purposes associated with most student ratings systems have tended to cause most students and/or faculty to produce scales that reflect the global characteristics of the "ideal teacher" and/or the "ideal class" with the expectations that when teachers are informed by their students that they and/or their classes deviate from this ideal, the teachers will make the adjustments of character, style, behavior or procedures that are required to match the model. Unfortunately, the relatively global descriptions that permitted scale creators to converge upon very similar descriptions of the ideal teacher-class, and that permit many scales to produce high inter-rater reliabilities, may not be very helpful to teachers that are informed that they rate low on "sense of humor," "seems to enjoy teaching," "seems to have self-confidence," "explains clearly," "knows if the class understands," "has a genuine interest in students," "has an interesting style of presentation" or "is valued for advice." Although many of these descriptors may be characteristic of many excellent teachers, they provide less than perfect direction to teachers who are not perceived by others to possess these characteristics to a high degree. Being informed that you are unprepared when you have spent hours in preparation--or that you are humorless, nonenergetic, unclear, lacking self-confidence and generally uninteresting--will not necessarily suggest strategies that you may employ to reduce these "defects." If ratings procedures and results are to be a primary guide to the improvement of teaching, ratings procedures should clearly describe potentially effective procedures and behaviors that can be substituted for those rated as ineffective.

Increasing the Feedback-Utility of Ratings

The IDEA system is in the process of constructing a series of short (8 to 20 pages) guidebooks that provide suggestions for teachers who are dissatisfied with their ratings. The various Pratical Guides for IDEA Users (e.g. Aspects of Lecturing by Brock, 1977) are keyed to low ratings of particular IDEA ratings items (e.g., "seemed enthusiastic about the subject matter," "made presentations that were dry and dull") and provide the teacher with a wide range of suggestions that are expected to match the styles of many different teachers. A similar series of short guidebooks (e.g. Practitioners' Views on Teaching the Large Introductory College Class, Brock, 1976) are less specifically tied to sets of ratings items but also offer a wide range of concise suggestions for the improvement of teaching.

Behaviorally Anchored Scales

An alternate approach to increasing the specificity of ratings has been described by Harari and Sedeck (1973) and Smith and Kendall (1963). Their suggested ratings forms consist of sets of descriptions of each major category of teaching activity. Each relatively independent teaching activity (e.g. Grading Procedures, Evaluation Procedures, Assignments, Course Organization, Presentation of Information, Stimulation, Student-Teacher Relationships, etc.) is described through the use of a set of decriptions of explicit behaviors (actual behavioral incidents) that have been ordered through consensus as reflecting different degrees of teaching quality. The resulting set of scales usually has as many pages as there are teaching functions, and each page contains a summary description of the teaching function, an ordered list of behaviors that characterize effective and ineffective teaching, and a point scale to which the behaviors are anchored. If the ordered list of behaviors are homogenous and also are ranked as the rater would rank them, and if a cluster of these behaviors actually reflect the teacher's typical behavior pattern, these scales can be used to produce a rating for each teaching function and sets of relatively independent ratings for each teacher. If the procedures of the scales are able to identify clearly described behaviors that characterize excellent teaching, instead of only identifying behaviors characteristic of poor teaching (evidently an easier task for most students), such scales might provide useful suggestions for the teacher who wishes to improve. In spite of the claim by Harari and Zedeck (1973) that such rating scales are easy for raters to use, the complexity of the decisions raters are asked to make, the forms' typically high reading load, the problems involved in printing the 8 to 10 page form, and the inability to optically-scan the student responses reduce the likelihood that these types of forms will replace the more typical ratings form for use with large groups of students and classes.

Low-inference Ratings Items

The replacement of high-inference descriptors with lower-inference items was explicitly attempted (Tom & Cushman, 1975) in the creation of the Cornell Diagnostic Observation and Reporting System for Student Description of College Teaching in order to provide more useful feedback to the instructors (to improve teaching). Although the resulting scale did include a series of relatively explicit descriptions of useful teaching behaviors (e.g., "gave step-by-step instructions when needed by students," "displayed concern that students learn," "spoke with expressiveness and variety in tone of voice," "praised students during class," "presented well-organized lectures," "addressed students by name," etc.), several of these "behaviors," especially when categorized as from "hardly ever" to "almost always," were far from perfectly clear. Such a compilation of odds and ends of frequently useful instructor actions ("used examples to make a point," "used understandable vocabulary," "stated the objectives of the course") may provide a compendium of useful advice, but one must wonder if the sum of these ratings is really that useful to the typical college teacher, and if the summary score is completely convincing as the most desireable, best indicator of teaching for use in personnel decisions. Although this type of scale is easier to print, use and process than the more complex but more coherent set of behaviorally anchored scales suggested by Harari and Zedeck (1973), it hardly seems superior to the IDEA scale (or similar scales) from which much of it was borrowed.

More Direct Advice From Students

Most ratings scales provide indications of deficits but fail to provide clear directions for change or clear suggestions for improvement. Although the open-ended questions associated with many ratings scales may elicit useful suggestions, they rarely provide clear guidance. Libbee (1976) has described a procedure more likely to provide teachers with guidance and useful suggestions. Teachers are asked to: (1) set aside one class hour in the middle of the term; (2) provide students with a short (5-7) set of open-ended questions (e.g., evaluate the course assignment and suggest improvements, evaluate the lectures and suggest improvements, etc.); (3) arrange the questions and answer spaces so that the pages of questions can be cut into slices that each contain a single question and its answer; and (4) divide the class into small groups so as to provide each group with a set of answers to one question. Each group of students is instructed to categorize these responses and rank order the categories as reflections of good teaching. The remaining twenty minutes left in the hour are used to present and discuss the results produced by each group. Libbee (1976) has written that these procedures can be shaped to reflect instructor concerns while providing students with a greater sense of participation and commitment to the course, and also with a greater sense of the diversity of student opinions. Wood (1977) described these procedures in a brief note to 154 college faculty. Several faculty expressed an interest in using the procedure, but only one high rated, very experienced teacher actually employed this procedure with his classes. It is possible that this type of procedure requires a degree of student-faculty rapport and teacher confidence that may not be available to those teachers that most need assistance and improvement.

The Georgia Teacher Assessment Project

The Teacher Assessment Program of the University of Georgia (Johnson, Okey, Capie, Ellett & Adams, 1978; Okey, Capie & Johnson, 1978) has developed a series of descriptions of generic teaching competencies and is validating observation scales related to these competencies. Although the focus of this project has been upon the teaching actions of pre-college teachers, many of these competency descriptions may be suitable for use with student or peer observations and ratings of college teaching performances. The descriptions employed in these "*Teacher Performance Assessment Instruments*" tend to represent a position midway between the typical student ratings scale and the behaviorally anchored scales described by Harari and Zedeck (1973) and may be useful for diagnosing classroom performance.

Some Faculty Characteristics

Information from ratings (or descriptions of the ideal teacher) more frequently has been demonstrated to be of use to pre-college teachers than to college teachers. Although the number of studies hardly justify firm conclusions, it is possible that college teachers are less easily changed or improved. A multitude of possibilities exist: (1) college teachers, either because of their characteristics, and/or those of their students or jobs, may already be closer to the ideal and therefore find it harder to take the last few steps; (2) the instruction associated with the shorter, more coherent, college course may be so dependent upon pre-planned course structures and syllabi that improvement is only likely if teachers are provided with a lead-time not provided by most investigators of the utility of ratings; (3) the college teacher is the American teacher least likely to have had the formal education concerning teaching methodology, educational evaluation and the nature, characteristics and development of their students that is required of most other teachers; or (4) the lack of supervised, practice teaching and the general lack of administrative supervision of college instruction has functioned to deprive many college teachers of any experience with supervision or guidance related to their teaching. The results drawn from a survey of college faculty (Cross, 1977; Cross, Mazur-Hart, & Kess, 1976) caused Cross (1977) to write that: (1) the level of faculty knowledge concerning recent educational information, research and innovation was low; (2) most faculty did not report that it was important to attend seminars on teaching or otherwise inform themselves about new knowledge in the area of teaching and learning; and (3) 94% of the faculty rated themselves as above average teachers, with 68% ranking themselves as in the top quarter of their faculty. The picture presented by Cross was of faculty who were not very knowledgeable concerning alternate approaches to instruction and who were not particularly motivated to learn more about alternate approaches to their own instruction. Nothing in the survey data produced by Cross, et al. (1976) would lead one to believe that the typically generous feedback provided by student raters was automatically likely to impel college teachers to improve their instruction--even if they knew how to improve their performance and/or their courses.

Incentives

The investigation by Roose (1977) and somewhat similar multivariate studies by Katz (1973), Kock and Chizmar (1973) and Siegfried and White (1973) have demonstrated that academic rank and years of experience are the major determinants of the salary of many faculty. However, while Roose (1977), Katz (1973) and Siegfried and White (1973) discovered that research, publication and/or administrative activity also were major predictors of absolute salary level, Kock and Chizmar (1973) discovered that rated merit, of which peer-rated teaching merit was the most important, explained over 71% of the variance in salary increments at an undergraduate-oriented state university. Siegfried and White (1973) studied a large and prestigious doctoral-oriented program and estimated that an improvement of one standard deviation in teacher evaluations would be expected to produce a salary increase of $490 (or 2.5%) while a similar increase in research productivity would earn $3,450, and even one article in a national journal would be worth over $390. Even the Kock and Chizmar (1973) study (at a former teachers college) indicated that a shift from being ranked as having only "some merit in teaching" up to the rank of "considerable merit in teaching" was worth only $15 per month (although the final increase to "unusual merit in teaching" would be worth $40 per month while a drop from "some merit" to "minimally acceptable" might cost $27 per month.) No other faculty activity was as likely to be associated with salary increments at this university. Two less empirically oriented studies of teaching-research trade-offs produced conflicting conclusions in that Shapiro (1978) tentatively concluded that a change in relative rewards would be successful in producing more teaching effort from a typical faculty (p. 65), while Hansen and Kelley (1973) concluded that "if the total salary budget to the faculty (e.g., a department) is fixed, then course evaluations will only lead to relatively small redistributions of income amongst the faculty" (p. 15). Hansen and Kelley (1973) also predicted that since, if total salary budgets are not changed, the introduction of a merit system associated with student (or other) evaluations of teaching will produce no total economic benefits for the faculty as a whole while creating costs for many faculty. Consequently, Hansen and Kenney (1973) predicted that, unless total benefits are increased, "it will be difficult to gain political support to implement and utilize course evaluations: rational behavior would predict notable resistance" (p. 15). Such resistance may underlie the inability of faculty to agree upon criteria or procedures for evaluating teaching effectiveness (see Cross, 1977; Guion, et al., 1977; or Wood, 1978 for examples of this disagreement).

The Hansen-Kelley approach to examining incentives, benefits and costs would be to ask, "Who is likely to benefit from the institution of student or peer evaluation of teaching (or advisement), and who is likely to lose?" In an institution in which research, publication, administrative activities, rank and length of service were the primary predictors of total salary and yearly increments, the addition of teaching merit as a component would most likely benefit those faculty who were confident that: (1) they were more able teachers than their departmental colleagues; (2) student and/or peer evaluations of teaching and/or advising would demonstrate this superiority; (3) the salary increment that they gained from the recognition of their teaching excellence would be greater than the salary increment that they lost when the traditional components of the salary increment (including an across-the-board percentage increase often allocated to attempt to offset inflation) were reduced in order to create the fund to be subdivided among the superior teachers. Quite obviously, the more prolific researchers, the administrators, and the more highly paid faculty would be more likely to have their previously-expected increments reduced, while only the lower ranking, less research interested faculty--who were also superior teachers--would be likely to benefit. Teachers who were interested in improving their teaching performance would not necessarily be helped by the institution of a department-wide or college-wide student ratings system since the lack of such a system hardly precludes their employment of the system of their own choice. Even administrators who are obligated to make personnel decisions might find that the questionable validity of ratings data and the controversy associated with the use of this data might not justify the increased information that ratings systems might provide. The generally unknown legal status of student ratings data might also deter some administrators from its gathering and employment in personnel decisions which might eventually be decided in court. In summation: (1) although many colleges appear to allocate segments of yearly salary increments to meritorious performance of different faculty functions; (2) increasing the proportion of the annual increment that is allocated to meritorious teaching performance may stimulate the improvement of some teachers; but (3) the benefits-cost balance associated with a shift in allocation toward teaching--especially teaching as evaluated by students or peers--may create notable faculty resistance. The lack of faculty agreement concerning the validity of most teacher evaluation systems may also create an additional resistance to the introduction and use of any new evaluation procedures.

SUGGESTIONS: A THREE PART PLAN

Part 1: Faculty Growth Contracting

Miller (1972) provided a detailed description of an individualized, evaluation and goal-setting procedure that has since been implemented in various forms within different departments and colleges (see Brett, 1978; Gross, 1976; Kramer, 1976; and Smith, 1976, pp. 60-67, for details of various "growth-contracting" plans). The faculty growth contracting model can be interpreted as an organized and systemized extention of the normal interaction between departmental faculty and their department chairperson. However, instead of the casual distribution of faculty assignments by the departmental administrator and the frantic gathering of evidences of merit by the faculty member being put forward for promotion or tenure, the contracting system encourages the faculty to regularly (usually annually, sometimes more often) establish a set of goals and objectives and submit these with plans for the new year to the departmental administrator--and/or to a team of colleagues. These development or growth plans are negotiated so as to coordinate individual and institutional resources and goals, and the activities included in the plan are continuously assessed according to criteria jointly established by faculty and their administrators and/or committees. Such a system includes a potential mentoring role for more experienced faculty and may initiate the development of various peer-support groups within (or beyond) the department. Growth contract systems may or may not be directly associated with summative or personnel evaluation and may or may not involve a "contract team" of colleagues that function as formal or informal, constructive evaluators and supporting guides.

Moomaw's (1977) survey and subsequent investigations of the faculty evaluation processes of southern colleges and universities led him to conclude that "the most glaring deficiency is that present approaches to evaluation do not seem to effectively stimulate and support faculty development and improvement even though institutions cite this as their primary evaluation goal" (pp. 89-90). The two chief reasons for this dissociation of evaluation and improvement are the lack of a link between evaluation and faculty development and the lack of faculty involvement in and control of the evaluation process (Moomaw, 1977, p. 90). The suggestions that concluded the Guion, et al. (1977) survey of faculty attitudes toward student ratings similarly focused upon the need for any faculty evaluation procedures to be employed to encourage improvement as well as to indicate current status. The faculty growth contract is one such system which clearly is focused upon the employment of evaluation to encourage faculty and departmental development. The growth contract also gives the individual faculty members a considerable control over and involvement with their own evaluation and assessment, and may even lead to the development of a network of supportive peer groups that can interact to forward one another's developmental efforts. Although Moomaw (1977) found growth contracting most often functioning in collegial and congenial, liberal arts colleges, some variant of such a system might be successful in a wide range of colleges.

Part II: A General Monitoring System

The creation of a growth contracting system would be unlikely to lead to the resolution of all problems associated with the formative and summative evaluation of faculty functions--although Moomaw (1977) wrote that, "in a growth contract institution the summative evaluation is neither so intimidating nor so surprising as it normally is because the faculty member is continually involved in evaluation, assessment, and growth and development...the individual is aware of his strengths and weaknesses...consequently, he and his immediate supervisor already know pretty much where he stands" (p. 89). Although individual faculty may wish to institute a wide range of assessment activities as a part of an individual growth contract, or to augment a personnel file (e.g., student ratings, peer visitations and evaluation of teaching performance; peer assessment of teaching materials and/or videotaped classes; peer assessment of course syllabi, examinations or of publications; evidences of service activities, etc.), at least a minimal, common monitoring procedure might be a useful supplement to individually initiated assessment activities. Most departments and colleges already maintain a relatively continuous record of the course, student and advisement loads of faculty. Regardless of the validity of student judgments concerning teaching effectiveness or advisement quality, declining enrollments, the potential relationships between student ratings and subsequent enrollments in similar courses (McKeachie & Solomon, 1958), and the possible relationship between student attrition rates and out-of-class contact with faculty (Rascarella & Terenzini, 1977), all appear to justify at least minimal consideration of student viewpoints. Such a minimal, but systematic assessment of student opinions would probably require: (1) the construction and employment of a short, global student ratings questionnaire with all classes; (2) the yearly assessment of student opinions of advisement procedures; and (3) an exit questionnaire (and/or interview to be completed by all students who graduate from (or otherwise leave) the department or its programs. The teaching ratings questionnaire need not contain more than a few questions designed to reflect general opinions concerning: (1) student opinions as to teaching effectiveness; (2) student assessments as to their own learning gains; (3) student beliefs that students were treated fairly and with courtesy and consideration; and (4) the effect of the course and/or the instructor's influence upon student interest in taking additional courses from the department. The short length of the form should permit its constructors to employ non-common sets of cues and adequately descriptive items while also requiring very limited instrusion into an instructor's limited class time. The item(s) reflecting instructor fairness might warn departmental administrators of impending problems while some (or all) of the other global items should be sufficient for personnel decision making. Those faculty that wished to augment the common questions with items more intended to provide diagnostic information would be free to do so whenever it suited their purposes. If these global items were employed by individual faculty or by departments for personnel purposes, the administration of the forms would have to be standardized and the information produced by students should probably be supplemented by more complete descriptions of the students and classes that produced the ratings (e.g., size and level of classes, average course grade). The use of global ratings for personnel purposes should also be supplemented by other evidences of teaching quality--especially by evidence reflecting those qualities that colleagues, rather than students, are more capable of evaluating (e.g., quality and appropriateness of course materials, objectives, syllabi and evaluation procedures and instruments).

Since the assessment of advising and the exit-questionnaire (or interviews) would be less frequently employed, one could justify the employment of longer, more detailed, and more diagnostically oriented instruments. Brock (1958) has described a questionnaire that has been developed to assess academic advisement, and individual institutions could employ this or locally developed advisement assessment instruments. The exit questionnaire would most likely have to be developed locally to reflect local conditions, goals and programs. Whether or not exiting students would be asked to rate or rank their experiences with courses, teachers and other aspects of the program or institution would depend upon the purposes and institution interests of the local departments.

The employment of general teaching and advisement monitoring within even a minimally operating growth contracting system would permit individual faculty and departmental administrators to adjust teaching and advisement assignments so as to attempt to produce a greater general level of teaching effectiveness and/or student (and teacher) satisfaction. If individual faculty contracts to weigh their advising and/or teaching heavily, and received poor global ratings, these teachers would still be able to make efforts to diagnose their difficulties and make improvements (or seek alternative indicators of effectiveness) before major personnel decisions (promotion, tenure, retention) had to be made. The use of a continuous monitoring system would provide faculty and administrators with sufficient advanced warning to permit modifications of assignments or practices before major personnel or programatic decisions had to be made. However, such a monitoring system would be ineffective unless its results were regularly shared by individual faculty and their administrator (or contracting mentor or committee).

If additional information must be gathered to provide students with better descriptions of courses, the need for such data could be identified through use of the exit questionnaires, and appropriate data could be gathered.

Part III: A Comprehensive Faculty/Instructional Development Support System

The faculty development center is a phenomenon that recently has blossomed on many college campuses (Centra, 1976; 1977a) and also in pre-college educational systems (Cauldwell, 1979; Hopgood, 1975; Julius, 1976). Berquist and Phillips (1975, 1977), Crow, et al. (1976), Diamond (1975), Linquist (1978), Moomaw (1977), and Smith (1976) have described several of the centers and activities that have been developed on many college campuses and the general theoretical models related to many of these activities. A useful, and perhaps necessary adjunct to a developmentally-oriented growth contracting system would be the availability of support services for those faculty that wished to improve one or more areas of faculty function. Most college faculty have had at least minimal training as researchers in their discipline, and many departments, colleges or universities already have facilities designed to support faculty research and grant-seeking. Consequently, calling the available resources to faculty attention may assist some faculty to begin to improve their research productivity. The major area that requires improvement at many colleges involves those evaluation, diagnostic, development and support services that may be required to assist those faculty wishing to improve their teaching, mentoring or advisement functions. Wood (1977c) has described several inexpensive procedures available to most college teachers who wish to improve their instruction, and the assessment and diagnostic forms produced by a wide variety of individuals and centers are widely available. Most university faculties contain experts in educational evaluation and test construction who are willing to share their expertise with their colleagues, and most also contain experts in public speaking, the use of media in instruction, and most other areas related to instructional improvement. Most universities also have a number of excellent teachers who represent a wide variety of teaching styles and strategies and who are able to function as mentors, models and/or guides for other faculty. The major impediments to faculty development are less likely to lie in the lack of facilities, assessment devices and models, and are more likely to be attributable to: (1) a failure to develop an adequate monitoring, supervision and support system capable of the early identification of potential problems; and (2) sensitive and nonthreatening guidance focused upon cooperation and improvement rather than upon the threat of external evaluation and punishment. Although there have been several recent surveys of instructional development activities on college campuses (Centra, 1976; Liebler, 1978; Moomaw, 1977), and other authors have attempted to determine program characteristics believed to lead to success (Lawrason & Hedberg, 1978; Wergin, 1977), the field is still new and still engaged in its own developmental and growing pains.

FI/S123

REFERENCE NOTES

1. Wood, P. H. *An analysis of student evaluations and self-reported characteristics of 500 graduate, teaching assistants*. Unpublished manuscript, 1978 (Available from the Office of Program Advisement, College of Education, BGSU, Bowling Green, Ohio 43403).

2. Wood, P. H. *An analysis of student ratings and course grades in a multi-section college course*. Unpublished manuscript, 1976 (Available from the Office of Program Advisement, College of Education, BGSU, Bowling Green, Ohio 43403).

3. Wood, P. H. *Student Description of Teaching: a student ratings form and its technical report*. Unpublished form and report, EDFI Department, BGSU, Bowling Green, Ohio, 1978.

4. Badia, P. Do methods of instruction and measures of different abilities and study habits interact? Unpublished manuscript, 1976. (Available from Dr. P. Badia, Department of Psychology, BGSU, Bowling Green, Ohio 43403).

5. Wood, P. H. & Bergman, G. The prediction of college grades through the use of a multifactor motivational index. Unpublished report, 1979. (Available from the Office of Program Advisement, College of Education, BGSU, Bowling Green, Ohio 43403).

sm2/9

REFERENCES

Aleamoni, L. M. The usefulness of student evaluations in improving college teaching. In A. L. Sockloff (Ed.), Proceedings: Faculty effectiveness as evaluated by students. Philadelphia: Measurement and Research Center, Temple University, 1973.

Aleamoni, L. M. The usefulness of student evaluations in improving college teaching (Research Report #356). Urbana, Illinois: The Office of Instructional Resources, University of Illinois, February, 1974.

Aleamoni, L. M. & Thomas, G. S. Is the instructor's rating of the class related to the class' rating of the instructor? Research Report No. 1. Tucson, Arizona: Office of Instructional Research and Development, University of Arizona, 1977.

Aleamoni, L. M., & Yimer, M. An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. Journal of Educational Psychology, 1973, 64, 274-277.

Alm, K. G., Ehrle, E. B., & Webster, B. R. Managing faculty reductions. Journal of Higher Education, 1977, 48, 153-163.

Arp, L., & Semb, G. An analysis of the use of student proctors in a personalized college business course. Journal of Personalized Instruction, 1977, 2(2), 92-95.

Astin, A. W., & Lee, C. B. T. Current practices in the evaluation and training of college teachers. In C. B. T. Lee (Ed.), Improving college teaching. Washington, D.C.: American Council on Education, 1967.

Ballard, M., Rearden, J., & Nelson, L. Student and peer rating of faculty. Teaching of Psychology, April 1976, 3, 88-90.

Batchelder, A. J., & Keane, F. An analysis of lecture in the college classroom through systematic observation. The Journal of Classroom Interaction, 1977, 13(1), 33-43.

Batista, E. E. The place of colleague evaluation in the appraisal of college teaching: A review of the literature. Research in Higher Education, 1976, 4, 257-271.

Bausell, R. B. Class size and student ratings: A methodological demonstration. An unpublished report. University of Delaware, Newark, Delaware, 197_ (no date).

Bausell, R. B., & Magoon, J. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement, 1972, 32, 1013-1023.

Bausell, R. B., Schwartz, S., & Purohit, A. An examination of the conditions under which various student rating parameters replicate across time. *Journal of Educational Measurement*, Winter 1975, *12*, 273-280.

Bausell, R. B., & Vinograd, C. J. Student ratings and various instructional variables from a within instructor perspective. Unpublished manuscript, 1977.

Bejar, I. I. A survey of selected administrative practices supporting student evaluation of instruction programs. *Research in Higher Education*, 1975, *3*, 77-86.

Bendig, A. W. A statistical report on a revision of the Miami Instructor Rating Sheet. *Journal of Educational Psychology*, 1952, *43*, 423-429.

Berquist, W. R., & Phillips, S. R., & Quehl, G. *A handbook for faculty development*. Washington, D.C.: Council for the Advancement of Small Colleges, Vol. 1, 1975; Vol. 2, 1977.

Blackburn, R. T., & Clark, M. An assessment of faculty performance: Some correlates between administrator, colleague, student and self ratings. Ann Arbor: Center for the Study of Higher Education, University of Michigan, 1971 (mimeographed).

Block, J. H., & Anderson, L. W. *Mastery learning in classroom instruction*. New York: MacMillan Publishing Company, 1975.

Bloom, B. S. *Human characteristics and school learning*. New York: McGraw-Hill, 1976.

Bloom, B. S. An introduction to mastery learning theory. In J. H. Block (Ed.), *Schools, Society and Mastery Learning*. New York: Holt, Rinehart and Winston, Inc., 1974.

Blum, M. L. An investigation of the relation existing between students' grades and their ratings of the instructor's ability to teach. *Journal of Educational Psychology*, 1936, *27*, 217-221.

Brandenburg, D. C., Derry, S., & Hengstler, D. D. Validation of an item classification scheme for a student rating item catalog. Paper presented at the NCME Annual Meeting, March 28, 1978, Toronto, Canada (available from the Office of Instructional Resources, University of Illinois, Urbana, Illinois 61801).

Braskamp, L. A. Student satisfaction as a useful criterion in departmental review. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March, 1978.

Braunstein, D. N., & Benston, G. J. Student and department chairman views of the performance of university professors. *Journal of Applied Psychology*, 1973, *58*, 244-249.

Braunstein, D. N., Klein, G. A., & Pachla, M. Feedback expectancy and shifts in student ratings of college faculty. Journal of Applied Psychology, 1973, 58, 254-258.

Breiter, J., & Menne, J. Measuring teacher performance. An unpublished and undated paper available from Dr. Joan Breiter or Dr. Jack Menne, Iowa State University.

Bresler, J. B. Teaching effectiveness and governmental awards. Science, 1968, 160, 164-168.

Brett, A. University of Miami's School of Education plan for faculty evaluation. Faculty Development & Evaluation in Higher Education, Spring 1978, 4(1), 16-18.

Brock, S. C. Aspects of lecturing: A practical guide for IDEA users. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, Kansas State University, 1977.

Brock, S. C. Measuring faculty advisor effectiveness. A paper presented at the Second Annual Conference on Academic Advisement, October 10, 1978. (Available from the Center for Faculty Evaluation and Development, Kansas State University, Manhattan, KS 66502.)

Brock, S. C. Practitioners' views on ... Teaching the large introductory college course. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, Kansas State University, 1976.

Brooks, D. W., & Levenson, H. Questions and answers about answers to questions. Journal of Chemical Education, 1974, 51, 161-162.

Bryan, R. C. Reactions to teachers by students, parents, and administrators. (United States Office of Education Cooperative Research Project No. 668). Kalamazoo, Michigan: Western Michigan University, 1963.

Burton, G. E., Pathak, D. S., & Burton, D. B. Effects of contract grading on student evaluation of teachers. Educational Horizons, 1977, 128-133.

Bussell, R. L., Costa, I. A., Spencer, R. E., & Aleamoni, L. M. Mermac Manual: Test and questionnaire analysis programs written for the IBM System/360. Urbana, Illinois: University of Illinois Press, 1971.

Cahen, L. S., & Filby, N. N. The class size/achievement issue: New evidence and a research plan. Phi Delta Kappan, March 1979, 60, 492-495, 538.

Caldwell, R. M. Transplanting the British teacher center in the U.S. Phi Delta Kappan, March 1979, 60, 517-520.

Calkins, D., Borich, G. D., Pascone, M., Kugle, C. L., & Marston, P. T. Generalizability of teacher behaviors across classroom observation systems. The Journal of Classroom Interaction, 1977, 13(1), 9-22.

Cartter, A. M. Ph.D.'s and the academic labor market. New York: McGraw-Hill Book Co., 1976.

Cashin, W. E., & Slawson, H. M. IDEA technical report No. 2: Description of data base, 1976-77. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, 1977a.

Cashin, W. E., & Slawson, H. M. IDEA technical report No. 3: Description of data base, 1977-78. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, 1977b.

Centra, J. A. Colleagues as raters of classroom instruction. Journal of Higher Education, 1975, 46, 327-337.

Centra, J. A. Faculty development practices. In J. A. Centra (Ed.), Renewing and Evaluating Teaching: New Directions for Higher Education, 1977(a), 1(1), 49-55.

Centra, J. A. Faculty development practices in U.S. colleges and universities (ETS PR-76-30). Princeton, N. J.: Educational Testing Service, November 1976.

Centra, J. A. The how and why of evaluating teaching. In J. A. Centra (Ed.), Renewing and Evaluating Teaching: New Directions for Higher Education, 1977(b), 1(1), 93-106.

Centra, J. A. How universities evaluate faculty performance: A survey of department heads--An abstract. In A. B. Smith (Ed.), Faculty Development & Evaluation in Higher Education: A Quarterly Newspaper, Fall, 1977(c), 3(3), 19.

Centra, J. A. The relationship between student and alumni ratings of teachers (Research Bulletin 73-39). Princeton, N. J.: Educational Testing Service, May 1973.

Centra, J. A. The student instructional report: Its development and uses (SIR Report #1). Princeton, N. J.: Educational Testing Service, 1972.

Centra, J. A. Student ratings of instruction and their relationship to student learning. In J. A. Centra (Ed.), Two studies on the validity of the Student Instructional Report (SIR Report No. 4). Princeton, New Jersey: Educational Testing Service, 1976.

Centra, J. A. The utility of student ratings for instructional improvement (Project Report 72-16). Princeton, New Jersey: Educational Testing Service, 1972.

Centra, J. A., & Creech, F. R. The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. In Two studies on the validity of the Student Instructional Report (SIR Report No. 4). Princeton, N. J.: Educational Testing Service, 1976, 17-74.

Clark, M. J. *Program review practices of university departments*. Graduate Record Examinations Board Research Report No. 75-5aR. Princeton, N. J.: Educational Testing Service, 1977.

Clark, D. L., & Guba, E. G. *RITE occasional papers series*. (Vol. 1-8). Bloomington, Indiana: Research on Institutions of Teacher Education, Indiana University, 1976.

Cohen, J., & Humphreys, L. G. Memorandum to faculty. Unpublished paper, University of Illinois, Department of Psychology, 1960.

Cole, C. C., Jr. *To improve instruction*. (ERIC/Higher Education Research Report No. 2). Washington, D.C.: The American Association of Higher Education, 1978.

Colliver, J. A. A report on student evaluation of faculty teaching performance at Sangamon State University. Technical Paper No. 1. Springfield, Illinois: Office of the Vice President, Sangamon State University, 1972.

Cooke, L. S. An analysis of certain factors which affect student attitudes toward a basic college course. Unpublished doctoral dissertation, Michigan State College, 1952.

Cope, R. G., McMillin, J. G., & Richardson, J. M. A study of the relationship between quality instruction as perceived by students and research productivity in academic departments. Washington: U. S. Office of Education, 1972.

Costin, F. Do student ratings of college teachers predict student achievement? *Teaching of Psychology*, April 1978, *5*, 86-88.

Costin, F. A graduate course in the teaching of psychology: Description and evaluation. *Journal of Teacher Education*, 1968, *19*, 425-432.

Costin, F. Intercorrelations between students' and course chairmen's ratings of instructors. University of Illinois, Division of General Studies, 1966. (mimeographed)

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, 1971, *41*, 511-535.

Crawford, P. L., & Bradshaw, H. L. Perception of characteristics of effective university teachers: A scaling analysis. *Educational and Psychological Measurement*, 1968, *28*, 1079-1085.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Cross, K. P. Not *can*, but *will* college teaching be improved? In J. A. Centra (Ed.), *Renewing and evaluating teaching: New Directions for Higher Education*, Spring, 1977, San Francisco: Jossey-Bass, Inc., pp. 1-15.

Cross, K. P., Mazur-Hart, S. F., & Kess, T. W. *Questionnaire for the faculty survey on teaching and learning*. Lincoln, Nebraska: University of Nebraska, 1976.

Crow, M. L., Milton, O., Moomaw, W. E., & O'Connell, W. R., Jr. *Faculty development in Southern Universities*. Atlanta: Southern Regional Education Board, 1976.

Daw, R. W., & Gage, N. L. Effect of feedback from teachers to principals. *Journal of Educational Psychology*, 1967, *58*, 181-188.

Delaney, E. L. The relationships of student ratings of instruction to student, instructor and course characteristics. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Delaney, E. L., Jr., & Kojaku, L. K. The influence of teaching experience and instructional development activities on student ratings of instruction obtained by beginning university professors. Paper read at the Annual Meeting of the American Educational Research Association, 1977.

Derry, J. O. Can students' ratings of instruction serve rival purposes? *Journal of Higher Education*, 1979, *50*(1), 79-88.

Derry, J. O. Strengths and Vulnerabilities of the Cafeteria Model. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.

Derry, J. O., Seibert, W. F., Starry, A. R., Van Horn, J. W., & Wright, G. L. The Cafeteria System: A New Approach to Course and Instructional Evaluation. (Report IRB 74-1). W. Lafayette, Indiana: Measurement and Research Center, Purdue University, 1974.

Diamond, R. M. *Syracuse University's Center for Instructional Development*. Syracuse, New York: Syracuse University, Center for Instructional Development, 1975.

Downie, N. M. Student evaluation of faculty. *Journal of Higher Education*, 1952, *23*, 495-496, 503.

Doyle, K. O. Faculty evaluation: Some considerations and a model. In A. L. Sockloff (Ed.), *Proceedings of the first invitational conference on faculty effectiveness as evaluated by students*. Philadelphia: Measurement and Research Center, Temple University, 1973, 59-90.

Doyle, K. O., Jr. *Student evaluation of instruction*. Lexington, Mass.: D. C. Heath & Company, 1975.

Doyle, W. Paradigms for research on teacher effectiveness. In L. S. Shulman (Ed.), *Review of research in education*. Itasca, Illinois: F. E. Peacock Publishers and the American Education Research Association, 1977.

Driscoll, L. A., & Goodwin, W. L. The effects of varying information about use and disposition of results on university students' evaluation of faculty and courses. *American Educational Research Journal*, Winter, 1979, *16*, 25-37.

Drucker, A. J., & Remmers, H. H. Do alumni and students differ in their attitudes toward instructors? *Journal of Educational Psychology*, 1951, *42*(3), 129-142.

Dunnette, M. D. Accuracy of students' reported honor point averages. *Journal of Applied Psychology*, 1952, *36*, 20-22.

Durbin, R., & Taveggia, T. C. *The Teaching-Learning paradox: A comparative analysis of college teaching methods*. Eugene, Oregon: Center for the Advanced Study of Educational Administration, 1968.

Elliot, D. H. Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education, 1950, *70*, 5-61.

Elmore, P. B., & LaPointe, K. A. Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology*, 1975, *67*, 368-374.

Elmore, P. B., & Pohlman, J. T. Effect of teacher, student and class characteristics on the evaluation of college instructors. *Journal of Educational Psychology*, 1978, *70*, 187-192.

Emmers, E. T., McBurnette, P., & Davis, O. L., Jr. Instructor perception, content of scale, and feedback effectiveness. Unpublished paper presented at the Annual Meeting of the American Educational Research Association, April, 1974.

Erickson, G. R., & Erickson, B. L. Improving college teaching. *Journal of Higher Education*, 1979, *50*, 670-683.

Erickson, G. R. & Sheehan, D. S. An evaluation of a teaching improvement process for university faculty. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976. (ERIC Document Reproduction Services No. ED 013 371).

Evans, E. D. Student activism and teaching effectiveness: Survival of the fittest? *Journal of College Student Personnel*, 1969, *10*, 102-108.

Faia, M. A. Teaching and research: Rapport or mesalliance. *Research in Higher Education*, 1976, *4*, 235-246.

Faricy, W. H. Department affiliation and faculty reference group: Some effects on faculty opinions. Chapter 2 or P. L. Dressel & W. H. Faricy, *Return to Responsibility*, San Francisco: Jossey-Bass, 1972.

Feild, H. S., Holley, W. H., & Armenakis, A. A. Effect of sex, class standing, and academic performance on perceived importance of teacher behaviors. *Research in Higher Education*, 1976, *5*, 215-222.

Feldman, K. A. Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education, 1977, 6, 223-274.

Feldman, K. A. Course characteristics and college students' ratings of their teachers: What we know and what we don't. Research in Higher Education, 1978, 9, 199-242.

Feldman, K. A. Grades and college students' evaluations of their courses and teachers. Research in Higher Education, 1976, 4, 69-111. (a)

Feldman, K. A. The significance of circumstances for college students' ratings of their teachers and courses. An article accepted for publication in Research in Higher Education, (1979?).

Feldman, K. A. The superior college teacher from the students' view. Research in Higher Education, 1976, 5, 243-288. (b)

Fenker, R. M. The evaluation of university faculty and administrators: A case study. Journal of Higher Education, November/December 1975, 46, 665-686.

Fentress, J. The effect of instructor influential tactics on their evaluation by university students. Unpublished Master's thesis, Bowling Green State University, Bowling Green, Ohio, 1973.

Fentress, J. H., & Swanson, R. A. The effect of instructor influential tactics on evaluation by university students. Journal of Industrial Teacher Education, 1975, 13(1), 5-16.

Flood Page, C. Student evaluation of teaching: The American experience. London: Soceity for Research into Higher Education, 1974.

Follman, J., Lavely, C., Silverman, S., & Merica, J. Student raters' referents in rating college teaching effectiveness. Journal of Psychology, 1974, 86, 247-249.

Follman, J., Lucoff, M., Small, L., & Power, F. Kinds of keys of student ratings of faculty teaching effectiveness. Research in Higher Education, 1974, 2, 173-179.

Follman, J., Merica, J. A., & Silverman, S. College students' ratings of trait names, definitions, descriptions, and combinations. Psychology, 1974a, 11, 11-12.

Follman, J., Merica, J. A., & Silverman, S. Negative numbers and order of number in student ratings. Perceptual and Motor Skills, 1974b, 38, 10.

Francis, J. B. Faculty ratings of course evaluation items. Research in Higher Education, 1976, 4, 23-40.

French, G. M. College students' concept of effective teaching determined by an analysis of teacher ratings. Dissertation Abstracts, 1957, 17, 1380-1381.

Frey, P. W. The Endeavor Instructional Rating Form. Evanston, Illinois: Department of Psychology, Northwestern University, 1973(a).

Frey, P. W. The Endeavor Instructional Ratings System: User's Handbook. Evanston, Illinois: Endeavor Information Systems (2407 Prospect Avenue, 60201), 1979.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85 (b).

Frey, P. W. A two-dimensional analysis of student ratings of instruction. Research in Higher Education, 1978, 9, 69-91.

Frey, P. W. Validity of student instructional ratings: Does timing matter? Journal of Higher Education, 1976, 47, 327-336.

Frey, P. W., Leonard, D. W., & Beatty, W. W. Student ratings of instruction: Validation research. American Educational Research Journal, 1975, 12, 435-447.

Frick, T., & Semmel, M. I. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 1978, 48, 157-184.

Fuller, F. F., & Manning, B. A. Teacher self-confrontation reviewed: A conceptualization for video playback in teacher education. Review of Educational Research, Fall 1973, 43, 469-528.

Gaff, J. G. Teachers who make a difference. In A. L. Sockloff (Ed.), Proceedings of first invitational conference on faculty effectiveness as evaluated by students. Philadelphia, Pennsylvania: The Measurement and Research Center of Temple University, April 1973, 219-232.

Gage, N. L. The appraisal of college teaching. Journal of Higher Education, 1961, 32, 17-22.

Gage, N. L. The scientific basis of the art of teaching. New York: Teachers College Press, 1978(a).

Gage, N. L. The yield of research on teaching. Phi Delta Kappan, November 1978, 60, 229-235.

Gage, N. L. Runkel, P. J., & Chatterjee, B. B. Changing teacher behavior through feedback from pupils: An application of equilibruim theory. In W. W. Charters & N. L. Gage (Eds.), Readings in the social psychology of education. Boston: Allyn & Bacon, 1963.

Gessner, P. K. Evaluation of instruction. Science, 1973, 180, 566-570.

Gillmore, G. M. A brief description of the University of Washington Instructional Assessment System. EAC Reports, Educational Assessment Center, University of Washington, 1974 (mimeo).

Gillmore, G. M. Statistical analysis of the data from the first year of use of the Student Rating Forms of the University of Washington Instructional Assessment System. EAC Report 76-9. Seattle, Washington: Educational Assessment Center, University of Washington, 1977.

Gillmore, G. M. Three functions of student course evaluations. Paper presented at a symposium on course evaluation, NCME Convention, April 6, 1972.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The teacher and the course as units of analysis in the generalizability of student ratings of instruction (EAC Report 77-9). Seattle, Washington: University of Washington Educational Assessment Center, February 1977.

Gillmore, G. M., & Naccarato, R. W. The effect of factors outside the instructor's control on student ratings of instruction. Seattle, Washington: Educational Assessment Center, University of Washington, 1975.

Glass, G. V. A paradox about excellence of schools and the people in them. Educational Researcher, March 1975, 4(3), 9-13.

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5(10), 3-8.

Glass, G. V., & Smith, M. L. Meta-Analysis of research on the relationship of class size and achievement. Boulder, Colorado: Laboratory of Educational Research, University of Colorado, September 1978.

Glass, G. V., Wilson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder, Colorado: Colorado Associated University Press, 1975.

Granzin, K. L., & Painter, J. J. A new explanation for students' course evaluation tendencies. American Educational Research Journal, 1973, 10, 115-124.

Grasha, A. F. The role of internal instructor frames of reference in the student rating process. Journal of Educational Psychology, 1975, 67, 451-460.

Gross, R. F. Faculty growth contracts. Faculty Development and Evaluation in Higher Education, Spring 1976, 2, 9-14.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Guion, R., Hutchinson, P., Klein, T., Statz, J., & Wood, P. Student evaluations of teaching: A report of the ad hoc committee to study the current uses of and faculty attitudes toward student evaluation of teachers and courses at Bowling Green State University (BGSU), 1977. (ERIC Document Reproduction Service No. ED 134 130).

Guthrie, E. R. The evaluation of teaching. *The Educational Record*, April 1949, *30*, 109-115.

Guthrie, E. R. *The evaluation of teaching: A progress report*. Seattle: The University of Washington, 1954.

Hammond, K., & Adelman, L. Science, values, and human judgment. *Science*, 1976, *194*, 389-396.

Hammond, K., Stewart, T., Brehmer, B., & Steinmann, D. In M. Kaplan and S. Schwarts (Eds.), *Human judgment and decision processes: Formal and mathematical approaches*. New York: Academic Press, 1975.

Hansen, W. L., & Kelley, A. C. Political economy of course evaluations. *The Journal of Economic Education*, Fall 1973, *5*(1), 10-21.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. *Journal of Applied Psychology*, 1973, *58*, 261-265.

Harnett, R. T., & Centra, J. A. *The effects of academic departments on student learning* (RB-76-19). Princeton, N.J.: Educational Testing Service, 1976.

Harvey, J. N., & Barker, D. G. Student evaluations of teaching effectiveness. *Improving College and University Teaching*, 1970, *18*, 275-278.

Haslett, B. J. Student knowledgeability, student sex, class size, and class level: Their interactions and influences on student ratings of instruction. *Research in Higher Education*, 1976, *5*, 39-65.

Hays, W. L. *Statistics*. New York: Holt, Rinehart and Winston, 1963.

Heilman, J. D., & Armentrout, W. D. The rating of college teachers on ten traits by their students. *Journal of Educational Psychology*, 1936, *27*, 197-216.

Hildebrand, M., Wilson, R. C., & Dienst, E. R. *Evaluating university teaching*. Berkeley, California: Center for Research and Development in Higher Education, University of California, 1971.

Hills, J. R. On the use of student ratings of faculty in determination of pay, promotion and tenure. *Research in Higher Education*, 1974, *2*, 317-324.

Hodgkinson, H. L., Hurst, J., & Levine, H. Current evaluation practices in "innovative" colleges and universities. *The Research Reporter* (The Center for Research and Development in Higher Education, University of California, Berkeley), 1975, *9*(1), 1-4.

Hogan, T. P. Similarity of student ratings across instructors, courses, and time. *Research in Higher Education*, 1973, *1*, 149-154.

Holland, J. B. The image of the instructor as it is related to class size. *Journal of Experimental Education*, 1954, *23*, 171-177.

Holmes, D. S. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, 1972, *63*, 130-133.

Hopgood, M. (Ed.) *Supporting the learning teacher: A source book for teacher centers*. New York: Agathon Press, 1975.

Hoyt, D. P. *Instructional Effectiveness: I. Measurement of Effectiveness; II. Identifying Effective Classroom Procedures*. (Research Reports No. 7, 8) Manhattan, Kansas: Office of Educational Research, Kansas State University, November, December, 1969.

Hoyt, D. P. Measurement of instructional effectiveness. *Research in Higher Education*, 1973, *1*, 367-378.

Hoyt, D. P., & Cashin, W. E. *IDEA technical report no. 1: Development of the IDEA system*. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, 1977.

Hoyt, D. P., & Howard, G. S. *The evaluation of faculty development programs* (Research Report #39). Manhattan, Kansas: Office of Educational Research, Kansas State University, May, 1977.

Hoyt, D. P., & Howard, G. S. The evaluation of faculty development programs. *Research in Higher Education*, 1978, *8*, 25-38.

Hoyt, D. P., & Spangler, R. K. Faculty research involvement and instructional outcomes. *Research in Higher Education*, 1976, *4*, 113-122.

Jaeger, R. M., & Freijo, T. D. Some psychometric questions in the evaluation of professors. *Journal of Educational Psychology*, 1974, *66*, 416-423.

Julius, A. British teacher centers: Practical applications for America. *Phi Delta Kappan*, November 1976, 251.

Kane, M. T., Crooks, T. J., & Gillmore, G. M. Generalizability and the interpretation of student evaluations of teaching. Paper presented at the Annual Meeting of the American Educational Research Association, 1976.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, 1976, *13*, 171-184.

Katz, D. A. Faculty salaries, promotions, and productivity at a large university. *American Economic Review*, June 1973, *63*, 469-477.

Keavney, G., & Sinclair, K. E. Teacher concerns and teacher anxiety: A neglected topic of classroom research. *Review of Educational Research*, 1978, *48*, 273-290.

Keller, F. S. "Good-bye teacher..." *Journal of Applied Behavioral Analysis*, 1968, *1*, 79-89.

Kemerer, F. R. The clouded future of faculty governance. *The Educational Forum*, January 1978, *43*, 233-243.

Koch, J. V., & Chizmar, J. F. The influence of teaching and other factors upon absolute salaries and salary increments at Illinois State University. *The Journal of Economic Education*, Fall 1973, *5*(1), 27-34.

Kramer, J. L. *Some suggestions for the management of human resources: Procedures at Camelot State College*. Manhattan, Kansas: Center for Faculty Evaluation and Development in Higher Education, Kansas State University, 1976.

Kulik, J. A., Kulik, C.-L. C., & Hertzler, E. C. Modular college teaching with and without required remediation. *Journal of Personalized Instruction*, 1977, *2*(2), 70-75.

Landis, L. M., & Pirro, E. B. Required/elective student differences in course evaluations. *Teaching Political Science*, 1977, *4*, 405-422.

Leftwich, W. H., & Remmers, H. H. A comparison of graphic and forced-choice ratings of teaching performance at the college and university level. *Purdue University Studies in Higher Education*, 1962, *92*, 3-31.

Lerner, B. Equal protection and external screening: Davis, DeFunis, and Bakke. *Proceedings of the 1977 ETS Invitational Conference*. Princeton, N. J.: Educational Testing Service, 1978(a).

Lerner, B. The Supreme Court and the APA, AERA, NCME test standards: Past references and future possibilities. Summary of the Spring 1978 Conference of the National Consortium on Testing. Cambridge, Mass.: Huron Institute, June 5, 1978.

Lester, R. *Antibias regulations of universities: Faculty problems and their solutions*. New York: McGraw-Hill Book Co., 1974.

Libbee, M. Use of students to diagnose teaching. *Faculty Development and Evaluation in Higher Education*, Fall 1976, *2*(3).

Lindquist, J. Dissemination projects in college teaching improvement. *Faculty Development and Evaluation in Higher Education*, Spring 1978, *4*(1), 18-20.

Love, J. A., & Sandoval, J. The generalizability and concurrent validity of an instrument to describe college teaching. A paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Love, J. A., Minnis, D. L., Sandoval, J., & Yonge, G. D. An instrument to describe college teaching: A reliability study. A paper presented at the Annual Meeting of the American Educational Research Association, New York, 1977.

Lovell, G. D., & Haner, C. F. Forced-choice applied to college faculty rating. Educational and psychological measurement, 1955, 15, 291-304.

Marsh, H. W. The relationship between students' evaluations of instructional quality (OIS 76-15). Los Angeles, California: Office of Institutional Studies, University of Southern California, 1976.

Marsh, H. W. Students' evaluation of instructional effectiveness: Relationship to student, course, and instructor characteristics. Paper read at the Annual Meeting of the American Educational Research Association, 1978.

Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14, 441-447.

Marsh, H. W., Fleiner, H., & Thomas, C. S. Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology, 1975, 67, 833-839.

Marsh, H. W., & Overall, J. U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and Variability Among College Students in Rating Their Teachers and Courses." Research in Higher Education, 1979, 10, 139-147.

Marsh, H. W., Overall, J. U., & Kesler, S. P. Class size, students' evaluations, and instructional effectiveness. American Educational Research Journal, Winter 1979, 16, 57-69.

Marsh, H. w., Overall, J. U., & Thomas, C. S. The relationship between student evaluations of instruction and expected grade. Paper read at the Annual Meeting of the American Educational Research Association, 1976.

Maslow, A. H., & Zimmerman, W. College teaching ability, scholarly activity and personality. The Journal of Educational Psychology, 1956, 47, 185-189.

Mayo, G. D. Faculty evaluation of a faculty development center. Research in Higher Education, 1979, 10, 253-262.

Mayr, E. Tenure: A sacred cow? Science, 24 March 1978, 199, editorial page.

McGrath, E. J. Characteristics of outstanding college teachers: Scholarship is the indispensable attribute. Journal of Higher Education, 1962, 33(2), 148-152.

McKeachie, W. J. New developments in teaching. New Dimensions in Higher Education, #16, Washington, D. C.: United States Office of Education, April, 1967.

McKeachie, W. J. Research on teahing at the college and university level. In N. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

McKeachie, W. J. Teaching tips. Lexington, Mass.: D. C. Heath & Company, 5th edition, 1965; 6th edition, 1969.

McKeachie, W. J, & Kulik, J. A. Effective college teaching. In F. N. Kerlinger (Ed.), Review of research in education (Vol. 3). Itasca, Illinois: F. E. Peacock, 1975.

McKeachie, W. J., & Lin, Y. Sex difference in student responses to college teachers: Teacher warmth and teacher sex. American Educational Research Journal, 1971, 8, 221-226.

McKeachie, W. J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal, 1971, 8, 435-445.

McKeachie, W. J., & Solomon, D. Student ratings of instructors: A validity study. Journal of Educational Research, January 1958, 51, 379-382.

McMillan, J. H. The effect of effort and feedback on the formation of student attitudes. American Educational Research Journal, Summer 1977, 14, 317-330.

Milhalo, W. E. Social psychological components of occupational choice. Unpublished doctoral dissertation, Bowling Green, Ohio: Bowling Green State University, 1977.

Mihalo, W. E., & Weber, W. C. Code manual for project data centered on the 1974-75 academic year. Bowling Green, Ohio: Professional Development Program in the Graduate College, Bowling Green State University, 1977.

Miller, M. T. Instructor attitudes toward, and their use of, student ratings of teachers. Journal of Educational Psychology, 1971, 62, 235-239.

Miller, R. I. Evaluating faculty performance. San Francisco: Jossey-Bass, 1972.

Mirus, R. Some implications of student evaluation of teaching. Journal of Economic Education, 1973, 5, 35-37.

Moomaw, W. E. Practices and problems in evaluating instruction. In J. A. Centra's (Ed.), Renewing and evaluating teaching (New Directions for Higher Education, #17). San Francisco: Jossey-Bass, Inc., Spring 1977.

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. The Journal of Educational Psychology, 1956, 47, 79-88.

Morstain, B. R., & Gaff, J. Student attitudes toward teaching improvement. *Faculty Development and Evaluation in Higher Education*, Fall, 1977, *3*(3), 1-6.

Morstain, B. R., & Kraft, R. E. Educational orientations of college students: A typological analysis. *Research in Higher Education*, 1979, *10*, 237-251.

Murray, H. G. The validity of student ratings of teaching ability. Paper presented at the Canadian Psychological Association Annual Meeting, Montreal, 1972.

Nachman, M., & Opochinsky, S. The effects of different teaching methods: A methodological study. *Journal of Educational Psychology*, 1958, *49*, 245-249.

Novick, M. R. The influence of the law on professional measurement standards. Proceedings of the 1977 ETS Invitational Conference. Princeton, N. J.: Educational Testing Service, 1978.

Orlandi, L. R., & Consalvo, R. W. Blaming the victim of evaluation. *The CEDR Quarterly*, Summer 1977, *10*(2), 3; 7.

Overturf, C. L., Jr., & Price, E. C. Student rating of faculty at St. Johns River Junior College. Administrative team report. Palatka, Florida: St. Johns River Junior College, 1966.

Pambookian, Hagop S. Initial level of student evaluation of instruction as a source of influence on instructor change after feedback. *Journal of Educational Psychology*, 1974, *66*, 52-56.

Pascarella, E. T., & Terenzini, P. T. Patterns of student-faculty informal interaction beyond the classroom and voluntary freshman attrition. In R. J. Silverman (Ed.), *Journal of Higher Education*, September/October 1977, *48*(5).

Pasen, R. M., Frey, P. W., Menges, R. J., & Rath, G. J. Different administrative directions and student ratings of instruction: Cognitive versus affective effects. *Research in Higher Education*, 1978, *9*, 161-167.

Perry, R. P., Niemi, R. R., & Jones, K. Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. *Journal of Educational Psychology*, 1974, *66*, 851-856.

Pohlman, J. T. A description of effective college teaching in five disciplines as measured by student ratings. *Research in Higher Education*, 1976, *4*, 335-346.

Pohlman, J. T. A multivariate analysis of selected class characteristics and student ratings of instruction. *Multivariate Behavioral Research*, 1975, *10*, 81-92.

Pohlman, J. T. Summary of research on the relationship between student characteristics and student evaluations of instruction at Southern Illinois University. Carbondale, Illinois: Counseling and Testing Center, Southern Illinois, University, 1972.

Pratt, M., & Pratt, T. A. E. C. A study of student-teacher grading interaction process. *Improving College and University Teaching*, 1976, *24*, 73-81.

Ranney, G. B. An experiment to investigate possible instructor-induced item response bias in the Purdue CAFETERIA instructional evaluation system. A paper presented at the Seventh Annual Southeastern American Institute for Decision Sciences Meeting, February, 1977 (available as IRB 77-3 from the Measurement and Research Center, Purdue University, West Lafayette, Indiana).

Remmers, H. H. The college professor as the student sees him. *Purdue University Studies of Higher Education*, 1929, *29*, 75.

Remmers, H. H., & Brandenburg, G. C. Experimental data on the Purdue Rating Scale for Instructors. *Educational Administration and Supervision*, 1927, *13*, 519-527.

Remmers, H. H., & Elliot, D. H. *Manual, the Purdue rating scale for instruction*. West Lafayette, Indiana: University of Purdue, 1948.

Remmers, H. H., Hadley, L., & Long, J. K. Learning, effort, and attitudes as affected by class size in beginning college engineering mathematics. *Purdue University Studies in Higher Education*, *No. 19*, 1932, *32*.

Remmers, H. H., Martin, F. D., & Elliot, D. N. Are students' ratings of instructors related to their grades? *Purdue University Studies in Higher Education*, 1949, *66*, 17-26.

Richards, J. M., & Lutz, S. W. Predicting student accomplishment in college from the ACT assessment. *Journal of Educational Measurement* 1968, *5*, 17-29.

Robin, A. L. Behavioral instruction in the college classroom. *Review of Educational Research*, 1976, *46*, 313-354.

Rodin, M., & Rodin, B. Student evaluations of teachers. *Science*, 29 September 1972, *177*, 1164-1166.

Ronning, R. R., & Walsh, U. R. Effects of student anonymity-nonanonymity on the factor structure of a teacher rating form. *Research in Higher Education*, 1977, *6*, 363-371.

Rood, H. J. Legal issues in faculty termination: An analysis based on recent court cases. *Journal of Higher Education*, 1977, *48*, 123-152.

Roose, J. E. The use of social judgment theory to study the extent of unfair sex bias in the salaries of faculty members. Unpublished doctoral dissertation, Bowling Green State University, August, 1977.

Searle, P. R. The effect of videotape feedback on the behavior of counseling groups. Dissertation Abstracts International, 1969, 29 (9-A), 2970.

Seiler, L. H., Weybright, L. D., & Stang, D. J. How useful are published evaluations ratings to students selecting courses and instructors? Teaching of Psychology, 1977, 4, 174-177.

Seldin, P., & Wakin, E. How deans evaluate teachers. Change, November 1974, 48-49.

Shapiro, E. The effects on teaching of changes in relative rewards for research and teaching. Research in Higher Education, 1978, 9, 43-67.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 1969, 22, 251-263.

Shavelson, R., & Dempsey-Atwood, N. Generalizability of measures of teaching behavior. Review of Educational Research, 1976, 46, 553-611.

Sherman, B., & Blackburn, R. T. Personal characteristics and teaching effectiveness of college faculty. Journal of Educational Psychology, 1975, 67, 124-131.

Shingles, R. D. Faculty ratings: Procedures for interpreting student evaluations. American Educational Research Journal, 1977, 14, 459-470.

Shipka, T. A. The crisis in peer review. NEA Advocate, April-May 1978, 6-7.

Shull, H. The university tenure "Problem." Science, 8 October 1976, 194, editorial page.

Siegfried, J. J., & White, K. J. Teaching and publishing as determinants of academic salaries. The Journal of Economic Education, 1973, 4, 90-99.

Simon, A., & Boyer, E. G. (Eds.). Mirrors for behavior. Philadelphia: Research for Better Schools, Inc., 1970.

Smart, J. C., & Montgomery, J. R. (Eds.) Examining departmental management. San Francisco: Jossey-Bass, Inc., 1976.

Smeltzer, C. H., & Harter, R. S. Comparison of anonymous and signed ratings of teachers. Educational Outlook, 1934, 8, 76-84.

Smith, A. B. Faculty development and evaluation in higher education (ERIC/Higher Education Research Report No. 8). Washington, D.C.: The American Association for Higher Education, 1976.

Smith, P. C., & Kendall, L. M. Retransaltion of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Smock, R., & Crooks, T. A plan for the comprehensive evaluation of college teaching. Journal of Higher Education, 1973, 44, 577-586.

Smock, R., & Crooks, T. Diagnosis and remediation of instructional problems without the use of standardized instruments. Paper read at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Sockloff, A. L. Instruments for student evaluation of faculty: Ideal and actual. In A. L. Sockloff (Ed.), Proceedings of the first invitational conference on faculty effectiveness as evaluated by students. Philadelphia: Measurement and Research Center, Temple University, 1973, 132-151.

Sockloff, A. L., & Deabler, V. T. The construction of the faculty and course evaluation instrument. Research Report. Philadelphia, Pa.: Temple University Testing Bureau, March 1971.

Spencer, R. E. The course evaluation questionnaire: Manual of interpretation. Urbana, Illinois: Office of Instructional Research, University of Illinois, 1965.

Spencer, R. E. A history of the development of the Illinois course evaluation questionnaire (Research Report #306). Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1969(a).

Spencer, R. E. Research report: Some dimensions of the Illinois Course Evaluation Questionnaire (Research Report #303). Urban, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1969(b).

Spencer, R. E., & Aleamoni, L. M. The Illinois Course Evaluation Questionnaire: Manual of interpretation. Urbana, Illinois: The University of Illinois at Urbana, 1968.

Stang, D. J., & Seiler, L. H. Improving College and University Teaching Yearbook 1977, 1977, 7-9.

Starry, A. R., Derry, J. O., & Wright, G. L. An automated instructor and course appraisal system. Educational Technology, May 1973, 13, 61-64.

Stone, E. F., Spool, M. D., & Rabinowitz, S. Effects of anonymity and retaliatory potential on student evaluations of faculty performance. Research in Higher Education, 1977, 6, 313-325.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Swanson, R. A., & Sisson, D. J. The development, evaluation, and utilization of a department faculty appraisal system. Journal of Industrial Education, 1971, 9(1), 64-79.

Thielens, W., Jr. Undergraduate definitions of learning from teachers. Sociology of Education, 1977, 50(July), 158-181.

Tom, F. K. T., & Cushman, H. R. The Cornell Diagnostic Observation and Reporting System for Student Description of College Teaching. Search: Agriculture, 1975, 5(8), 1-27 (available from the Cornell University Agricultural Experiment Station, Cornell University, Ithaca, New York).

Touq, M. S., Feldhusen, J. F., & Halstead, J. Criterion-referenced validity of student ratings of instructors. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Truax, C. B., & Mitchell, K. M. Research on certain therapist interpersonal skills in reaction to process and outcome. In A. E. Bergin & S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change: An empirical analysis. New York: John Wiley, 1971.

Tuckman, B. W. A system for helping teachers to change their affective behavior through feedback. In A. L. Sockloff (Ed.), Faculty effectiveness as evaluated by students. Philadelphia: Measurement and Research Center, Temple University, 1973.

Tuckman, B. W., & Oliver, W. F. Effectiveness of feedback to teachers as a function of source. Journal of Educational Psychology, 1968, 59, 297-301.

Tukey, J. W. Some thoughts on clinical trials, especially problems of multiplicity. Science, 18 November, 1977, 198, 679-684.

Vecchio, R., & Costin, F. Predicting teacher effectiveness from graduate admissions predictors. American Educational Research Journal, Spring 1977, 14, 169-176.

Velicer, W. F. Expected grade and student ratings of instructors: A reappraisal. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.

Villano, M. W. The development and validation of an omnibus student rating form for evaluation of courses and instructors at a large university. Unpublished paper presented at the Annual Meeting of the National Council on Measurement in Education, April 5, 1977, New York.

Villano, M. W., Rosenstock, E. H., & Estes, C. "A decade with a student course evaluation form at a major university." Paper read at the Annual Meeting of the American Education Research Association, Chicago, Illinois, April, 1974.

Voecks, V. W., & French, G. M. Are student-ratings of teachers affected by grades? *Journal of Higher Education*, 1960, *31*, 330-334.

Wagner, A. C. Changing teaching behavior: A comparison of microteaching and cogntive discrimination training. *Journal of Educational Psychology*, 1973, *64*, 299-305.

Walker, B. D. An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers. *Dissertation Abstracts*, 1969, *29*(9-B), 3474. Ann Arbor, Michigan: University Microfilms, Order No. 69-786.

Wallen, N., & Travers, R. Analysis and investigation of teaching methods. In N. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Walsh, W. B. Validity of self-report. *Journal of Counseling Psychology*, 1967, *14*, 18-23.

Webb, W. B., & Nolan, C. Y. Student, Supervisor, and self-ratings of instructional proficiency. *The Journal of Educational Psychology*, 1955, *46*, 42-46.

Whitely, S. E., & Doyle, K. D. Implicit theories in student ratings. *American Educational Research Journal*, Fall 1976, *13*, 241-253.

Widlak, F. W., McDaniel, E. D., & Feldhusen. Factor analysis of an instructor rating scale. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February 27, 1973.

Wilson, R. C., Dienst, E. R., & Watson, N. L. Characteristics of effective college teachers as perceived by their colleagues. *Journal of Educational Measurement*, 1973, *10*(1), 31-37.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Wood, K., Linsky, A., & Straus, M. A. Class size and student evaluations of faculty. *Journal of Higher Education*, 1974, *45*, 524-534.

Wood, P. H. *The description and evaluation of a college department's faculty rating system*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1977. (ERIC Document Reproduction Service No. ED 142 128) (a).

Wood, P. H. Faculty attitudes toward student evaluation of teaching. *Faculty Development & Evaluation in Higher Education*, Summer 1978, *4*(2), 10-14.

Wood, P. H. A simple diagnostic teaching-evaluation technique. *Brief notes: Faculty development & evaluation news*, February 1977, *1*(1), 1-2. (Available from the EDFI Department, College of Education, BGSU, Bowling Green, Ohio 43403) (b).

Wood, P. H. *Student evaluation of teaching*. Bowling Green, Ohio: Bowling Green State University, June 1977 (ERIC Document Reproduction Service No. ED 134 130) (c).

Wood, P. *Student and peer rating of college teaching and peer rating of research and service: Four years of departmental evaluation*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978 (ERIC Document Reproduction Service No. ED 155 218).

Wood, P. H. *Teacher evaluation: Feedback to improve teaching*. Bowling Green, Ohio: Bowling Green State University, 1977. (ERIC Document Reproduction Service No. ED 139 733) (d).

sm2/10a&b